UNIVERSIDADE FEDERAL DE SANTA CATARINA

Reconhecimento automático de emoções através da voz

Jair da Rosa Júnior

Florianópolis

2017/2

Jair da Rosa Júnior

Reconhecimento automático de emoções através da voz

Trabalho de conclusão de curso apresentado como parte dos requisitos para obtenção do título de Bacharel, do curso de Sistemas de Informação na Universidade Federal de Santa Catarina.
Orientador: Profº Elder Rizzon Santos, Dr.

Florianópolis

2017/2

Jair da Rosa Júnior

Reconhecimento automático de emoções através da voz

Trabalho de conclusão de curso apresentado como parte dos requisitos para obtenção do título de Bacharel, do curso de Sistemas de Informação na Universidade Federal de Santa Catarina.

Profº Cristian Koliver, Dr.
Coordenador do Curso

Banca Examinadora:

Profº Elder Rizzon Santos, Dr.
Orientador
Universidade Federal de Santa Catarina

Profª Jerusa Marchi, Drª
Universidade Federal de Santa Catarina

Profº Mauro Roisenberg, Dr.
Universidade Federal de Santa Catarina

Thiago Angelo Gelaim, Me.
Universidade Federal de Santa Catarina

Florianópolis
2017/2

Agradecimentos

Agradeço primeiramente aos meus pais Jair e Tânia, e meus irmãos, por terem me incentivado e dado apoio para eu chegar até aqui, sem eles esta caminhada teria sido imensamente mais difícil.

À minha amada namorada e melhor amiga, que cooperou com o desenvolvimento deste trabalho, suportando amorosamente a minha ausência com total compreensão, permitindo que eu me dedicasse ao desenvolvimento deste trabalho.

Agradeço à todos os professores da UFSC que me ensinaram, de uma forma ou de outra, como superar os obstáculos que encontrei durante a caminhada nesta universidade, em especial ao meu orientador Profº Dr. Elder Rizzon Santos, que me ajudou imensamente, não somente com este trabalho de conclusão, mas também durante vários momentos da graduação.

Por fim, agradeço também aos membros da banca Profª Drª Jerusa Marchi, Profº Dr. Mauro Roisenberg e Me. Thiago Angelo Gelaim por terem aceitado compor a mesma.

Resumo

Após o surgimento dos telefones, e mais recentemente dos computadores, se tornou possível o armazenamento de áudios no formato digital. Os celulares modernos juntamente com a internet tornaram viável a gravação e transmissão destes áudios em larga escala. Surge então uma nova demanda de processamento e extração de informação dos mesmos. O reconhecimento de emoções através da voz é uma demanda recente, que só apareceu com a popularização de algoritmos de aprendizado de máquina, onde se destacam KNN, SVM, GMM e HMM. Neste trabalho foi proposto um sistema baseado em SVM, onde são extraídas características da voz (tais como pitch e energia) e um modelo é treinado de forma supervisionada, utilizando cada emoção a ser reconhecida como uma classe. O reconhecimento se dá, pela classe com maior verossimilhança obtida. Utilizando o banco de dados emocional de Berlin (em alemão) conseguimos obter uma taxa de reconhecimento de 86,79% e com o banco de dados criado em português, extraindo-se trechos de filmes e vídeos, foi obtida uma taxa de 70,83%. Os resultados obtidos foram bastante razoáveis, visto que alguns autores do estado da arte obtiveram resultados piores.

Palavras-chave: reconhecimento, emoção, SVM, KNN, GMM, HMM, voz, aprendizado de máquina, Emo-DB, banco de dados em português.

Abstract

After the emergence of the phones and more recently the computers, it become possible storing audios in digital format. Modern cell phones along with the internet, have made recording and transmitting these audios on a large scale viable. Then, a new demand of processing and extracting information arises. The speech emotion recognition is a recent demand, which only appeared because of the popularization of machine learning algorithms, where stands out KNN, SVM, GMM and HMM. In this work, we propose a SVM-based system, where voice features are extracted (like energy and pitch) and a supervised model is trained, utilizing each emotion to be recognized as a class. The recognition is given by the class with the highest likelihood. Using the Berlin Database of Emotional Speech (Emo-DB), we achieve a recognizing rate of 86,79% and using a Portuguese database, we've reached a rate of 70,83%. The obtained results were very reasonable, since some authors of the state-of-the-art got worse results.

Keywords: recognizing, emotion, SVM, KNN, GMM, HMM, speech, machine learning, Emo-DB, Portuguese Database;

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

CTI	*Computer Telephony Integration*
IA	Inteligência Artificial
ML	*Machine Learning*
SVM	*Support Vector Machine*
HMM	*Hidden Markov Model*
KNN	*K-Nearest Neighbors*
GMM	*Gaussian Mixture Model*
MFCC	*Mel Frequency Cepstral Coefficients*
LFPC	*Log Frequency Power Coefficients*
HNR	*Harmonic to Noise Ratio*
DSCC	*Delta-Spectral Cepstral Coefficients*
T-DSCC	*Teager based DSCC*
LPCC	*Linear Prediction Cepstral Coefficients*
LFPC	*Log Frequency Power Coefficients*
MFCC	*Mel Frequency Cepstral Coefficient*
CHMM	*Continous Hidden Markov Model*
SBC	*Sub-Band Cepstrum*
LDA	*Linear Discriminant Analysis*
NFD-LFPC	*nonlinear frequency domain LFPC*
NTD-LFPC	*nonlinear time domain LFPC*
TEO	*Teager Energy Operator*

DBN	*Deep Belief Network*
RBM	*Restricted Boltzmann Machines*
RNA	Rede Neural Artificial
HTK	*Hidden Markov Toolkit*
UAR	*Unweighted Average Recall*
EMO-DB	*Berlin Emotional Database of Speech*
AMDF	*Average Magnitude Difference Function*
BD	*Banco de Dados*
COPOM	*Centro de Operações da Polícia Militar*

Sumário

1 Introdução

A transmissão de voz por sistemas eletrônicos foi uma revolução inimaginável na forma como nos comunicamos e já existe há mais de um século. Com a popularização dos telefones e mais recentemente com a dos computadores, a integração entre telefonia e computação se tornou algo que certamente iria acontecer.

Hoje existem muitas dessas soluções de CTI (do inglês, *Computer Telephony Integration*) que, por exemplo, realizam gravações dos áudios de conversas telefônicas que podem ser utilizadas para diversos fins, desde somente para registro da conversa, como para realizar algum processamento sobre esta gravação. Os processamentos destas gravações podem ser diversos, mas um que gostaríamos de destacar é o reconhecimento de emoções.

Essencialmente, a fala serve para transmitir uma mensagem através de palavras. Contudo, ela pode transmitir muito mais do que apenas palavras, pois possui características intrínsecas à ela – sonoridade, passo, entonação, nitidez, articulação, irregularidade, instabilidade e velocidade de fala são algumas delas. Com a análise destas características por algoritmos de Inteligência Artificial (IA) torna-se possível o reconhecimento automático da emoção do interlocutor.

IA é um conceito que não tem uma definição exata. Pelo fato da inteligência em si não ser um conceito bem definido, vários autores conceituam o termo de diferentes formas. Segundo Luger (2013), a IA pode ser definida como o ramo da ciência da computação que se ocupa da automação do comportamento humano.

Dificilmente seria possível realizar o reconhecimento de emoções sem os algoritmos modernos de machine learning (ML). ML é um campo da área de inteligência artificial, onde são desenvolvidos algoritmos e técnicas que permitam às máquinas aprenderem. Certamente este aprendizado de máquina não funciona da mesma forma como nós, humanos, aprendemos. Luger (2013) afirma que o aprendizado envolve a generalização a partir da experiência, e é isto que um algoritmo eficiente de ML deve fazer: analisar um conjunto de dados amostrais e gerar um modelo que induza qual o resultado para amostras totalmente desconhecidas. Em ML, no contexto deste trabalho, preocupa-se com o tipo de raciocínio

indutivo, onde são analisados grandes conjuntos de dados para extração de alguns padrões e regras.

Alva, Nachamai e Paulose (2015) descrevem que computação emocional é uma área da inteligência artificial que busca preencher o gap entre emoções humanas e tecnologia da computação. Por estudar emoções, logicamente, este campo de pesquisa é multidisciplinar, envolvendo estudos das áreas de ciência da computação, ciência cognitiva e principalmente da psicologia.

Talvez isso passe despercebido no nosso dia-a-dia, mas as emoções guiam a maneira de percepção do mundo ao nosso redor. Logo, uma máquina que consegue compreender as emoções humanas poderá dar uma resposta mais adequada à emoção identificada no momento.

1.1 Motivação

O principal motivo para o desenvolvimento deste projeto é o fato de ainda não existir um sistema de reconhecimento de emoções que funcione com uma precisão aceitável e de forma genérica. Um sistema com tal precisão, pode ser bastante útil para reconhecer emoções em ligações telefônicas, em um *call center* por exemplo, visando identificar se um cliente está estressado ou nervoso com quem está lhe atendendo, podendo indicar a qualidade do atendimento.

Outro modo útil de aplicação seria, o reconhecimento de emoções durante um atendimento da central de emergência da polícia, onde a emoção da pessoa atendida poderia auxiliar em uma tomada de decisão, definição de prioridade de atendimento, e até mesmo identificação de trotes. Uma função interessante a ser citada, é a possibilidade de utilização do reconhecimento das emoções para aprimoramento da interação entre seres humanos e máquinas. Neste contexto, se um robô souber a emoção do humano com o qual está interagindo, pode dar uma resposta que melhor se adapta ao momento, e não somente mais uma resposta “padrão” como acontece atualmente.

1.2 Objetivos

1.2.1 Objetivo Geral

Este projeto tem como objetivo principal desenvolver um sistema que utiliza técnicas e algoritmos de *machine learning* para a identificação de padrões na voz através das características da mesma, com o intuito de realizar o reconhecimento automático de emoções. O sistema deve reconhecer a emoção representada por um áudio entre um grupo de emoções pré-definidas as quais o algoritmo foi treinado, ou seja, se trata de um problema de classificação.

1.2.2 Objetivos Específicos

Para o desenvolvimento do sistema e escolhas das tecnologias e técnicas a ser utilizadas, é necessário o estudo mais aprofundado sobre o estado da arte, e as técnicas e ferramentas disponíveis. Abaixo será apresentado alguns objetivos específicos para a conclusão do desenvolvimento do sistema proposto:

- Realizar pesquisa teórica sobre o estado da arte em ML para embasamento teórico sobre o assunto.;
- Analisar técnicas e algoritmos de ML, visando obter uma visão geral dos algoritmos existentes atualmente;
- Realizar testes com alguns dos algoritmos estudados no item 2, objetivando encontrar o mais apropriado para resolução do problema proposto;
- Avaliar ferramentas disponíveis que possam facilitar o estudo e desenvolvimento do sistema;
- Realizar os testes iniciais utilizando um banco de dados já consolidado para tornar possível a comparação com outros autores;
- Desenvolver o sistema proposto para reconhecimento de emoções também para a língua portuguesa;
- Definir as limitações do sistema que será desenvolvido, tais como o conjunto de emoções reconhecidas e de que forma se dará o reconhecimento;

- O sistema desenvolvido neste trabalho deve ser independente de gênero, pois a aplicação de um classificador de gênero antes do reconhecimento da emoção poderia aumentar muito o tempo de resposta;
- Testar e ajustar o sistema, propondo aprimoramentos para futuras pesquisas/trabalhos.

1.3 Método de Pesquisa

Por serem pré-requisitos para o entendimento dos algoritmos e teorias relacionadas ao desenvolvimento do sistema proposto, a etapa inicial do trabalho será composta pela pesquisa do estado da arte em IA e ML. Este estudo será realizado utilizando, principalmente, artigos *surveys*, teses, publicações e livros pertinentes ao assunto. A relevância das obras encontradas será definida em conjunto com o orientador.

Através do conhecimento adquirido na pesquisa, será desenvolvido o sistema proposto, fundamentado nos conceitos de IA, ajustando-se os parâmetros e dados (áudios) de treino para o sistema de ML. Após isto o sistema deve ser testado e os resultados esperados e os obtidos serão documentados. A análise dos resultados obtidos será o principal parâmetro para mensurar a qualidade do sistema desenvolvido.

1.4 Estrutura do Trabalho

Este trabalho foi dividido em 5 capítulos. Este capítulo introdutório descreve o contexto, motivação, metodologia e objetivos do presente trabalho. No capítulo seguinte são apresentados os principais fundamentos teóricos necessários para o entendimento do trabalho. O terceiro capítulo descreve alguns trabalhos relacionados com a proposta desta monografia, que são mais relevantes. A descrição do sistema proposto, banco de dados de áudio utilizados, decisões tomadas e resultados são apresentados nos capítulos finais. As conclusões são apresentadas no sexto capítulo, juntamente com os trabalhos futuros.

2 Fundamentação Teórica

Neste capítulo são apresentados conceitos relacionados à emoções (emoções existentes, modelos de emoções, etc.), reconhecimento através da voz, características (características existentes, extração, etc.) e também sobre os métodos de classificação (SVM, HMM, KNN e GMM). Os conceitos discutidos neste capítulo são essenciais para o entendimento e desenvolvimento do sistema que será proposto no capítulo 4.

2.1 Emoções

Emoções fazem parte do cotidiano, porém poucas pessoas sabem a fundo o que significa a palavra emoção. A fim de compreender o que é uma emoção, é necessário o estudo da teoria das emoções.

2.1.1 Teoria das Emoções

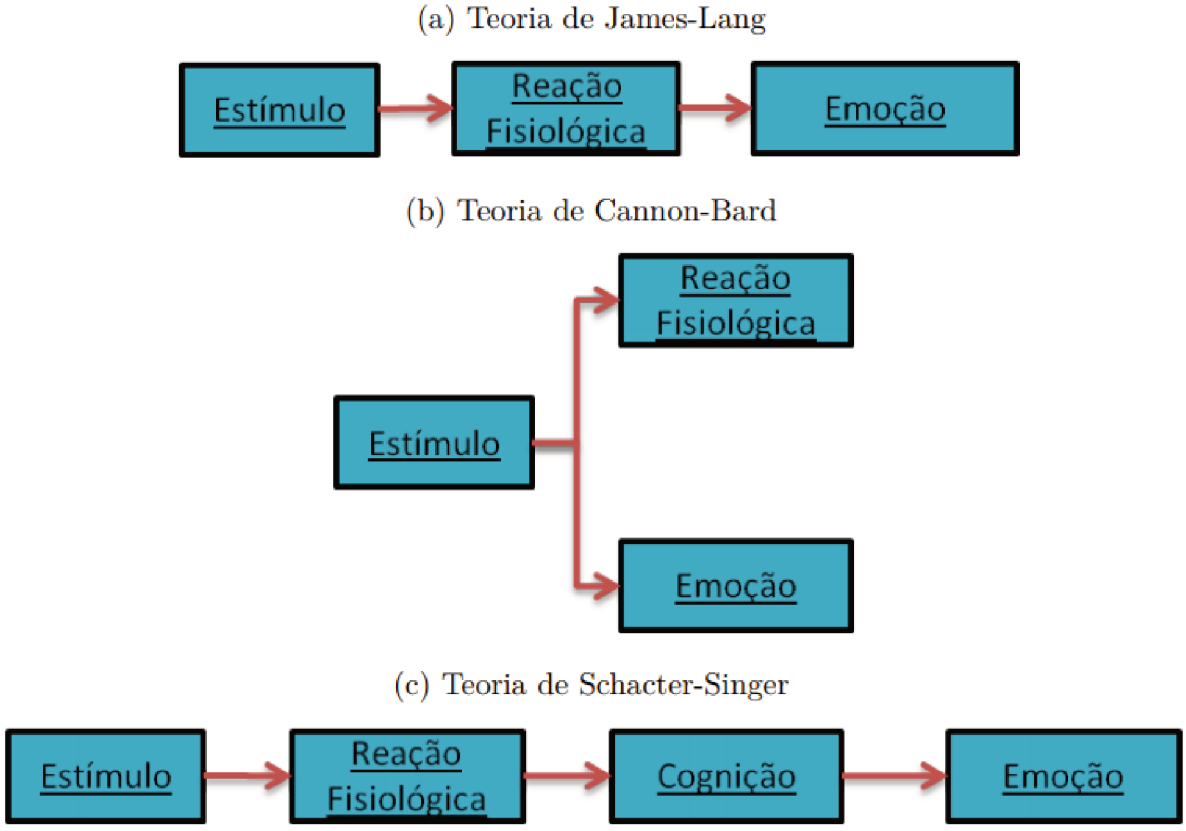
É necessário entender que o significado da palavra emoção é bastante amplo e, muitas vezes, é confundido com as reações do corpo ao vivenciar um estado emotivo. Estas reações são de suma importância para o reconhecimento automático de emoções, visto que esta tarefa consiste em avaliar as alterações na voz para concluir qual o estado emocional de um indivíduo.

Segundo Le e Provost (2013), a expressão da emoção é um processo dinâmico e complexo, portanto, o estudo das emoções envolve várias áreas, mas principalmente a psicologia, onde existem várias teorias que explicam a manifestação da emoção (HOUWER; HERMANS, 2010 apud IRIYA, 2014). As teorias de emoções mais conhecidas são a de James-Lang e de Cannon-Bard.

James-Lang propôs que um indivíduo, após receber um estímulo exterior, sofre alterações fisiológicas perturbadoras, sendo o reconhecimento desses sintomas pelo cérebro o que gera a emoção. Já a Teoria de Cannon-Bard sugere que as reações fisiológicas e a emoção ocorrem simultâneamente, uma vez que a interpretação do estímulo exterior ocorre em duas partes diferentes do cérebro. Há ainda algumas outras teorias, como as

cognitivistas, que afirmam que os processos cognitivos, como percepções e recordações, são fundamentais para se perceberem as emoções. Tendo como exemplo a teoria de Schachter-Singer, que presume que a experiência da emoção cresce a partir da consciência de excitação fisiológica (IRIYA, 2014).

Figura 1 Diferentes percepções das Teorias de Emoções



Fonte: Iriya (2014)

Existem também algumas outras teorias de emoções citadas no *survey* realizado por Gunes et al. (2011).

Estas teorias são distintas, porém todas afirmam que emoções possuem uma causa externa ao corpo humano, que geram alterações fisiológicas e que podem ser diretamente observadas. Estas afirmações são de vital importância para o desenvolvimento de um sistema de reconhecimento automático de emoções.

2.1.1.1 Emoções Existentes

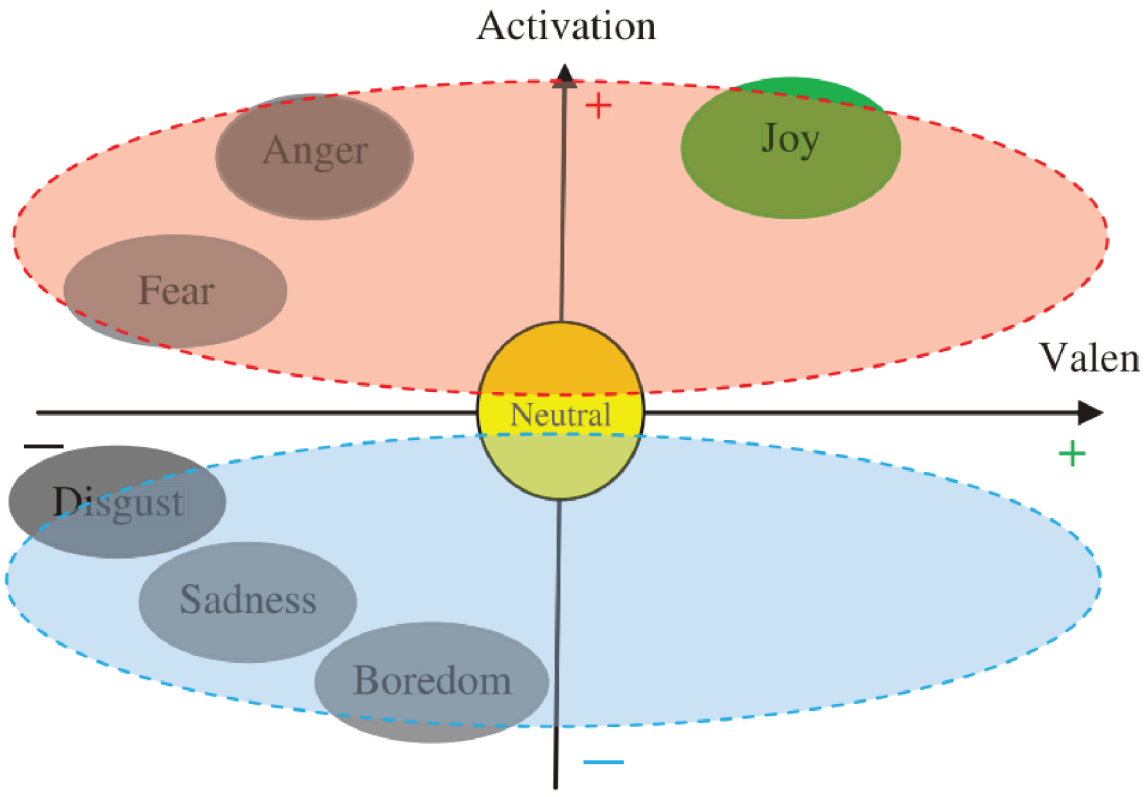
Com todas estas teorias, deve ser pensado em quais emoções existem, e quais devem ser reconhecidas. Um estudo realizado por Whissell et al. (1986) afirma que há

mais de 4 mil palavras que representam emoções, porém como quanto maior o número de emoções que o sistema reconhece mais sua precisão diminui, pois segundo Iriya (2014), deve-se definir um conjunto mais limitado possível somente com as emoções mais básicas, que sejam de fácil identificação do ponto de vista humano, e principalmente àquelas cujo reconhecimento automático tenha aplicação prática.

2.1.1.2 Modelo Dimensional

Existem vários modelos de emoções com duas ou três dimensões. Harimi et al. (2015) utilizou em seu trabalho apenas duas dimensões: ativação e valência, onde as posicionou em um plano bidimensional, conforme é exibido na figura 2.

Figura 2 Espaço de emoções bidimensional entre as dimensões ativação e valência.



Fonte: Harimi et al. (2015)

Iriya (2014) descreve três dimensões: Ativação, Avaliação e Domínio. A Ativação representa o grau de intensidade da emoção, ou seja, o quanto esta emoção altera o estado emocional do indivíduo, em comparação com o estado neutro. A Avaliação têm relação com a nossa noção de julgamento das emoções em positivas e negativas, boas e más. Esta última dimensão alguns modelos omitem, o Domínio, que tem a ver com o quanto a emoção

têm influência em um indivíduo, sendo o quanto ela o controla ou quanto a pessoa tem controle sobre a emoção.

> Embora a classificação baseada nestas três dimensões seja por vezes empírica ou baseada em senso comum, é possível que alguns aspectos sejam refletidos em reações fisiológicas, que podem ser medidas e utilizadas na análise como parâmetros fisiológicos. Por exemplo, é de se esperar que alguém em alto estado de excitação fale mais alto (IRIYA, 2014, p. 33).

Com esta afirmação, percebe-se que algumas emoções apresentam alterações fisiológicas mais perceptíveis do que outras, porém nenhum autor evidenciou que isto torne estas emoções mais fáceis de serem reconhecidas devido à este fator.

2.2 Reconhecimento

Definido o conjunto de emoções e os modelos de classificações, pode-se expor o que exatamente é esta importante tarefa de reconhecimento automático de emoções, que será explicada a seguir. Alguns aspectos relevantes sobre este tema serão abordados mais adiante nesta seção.

2.2.1 Reconhecimento Automático Através da Voz

O reconhecimento automático de emoções através da voz é realizado processando-se um arquivo de áudio, identificando as alterações da voz contidos nos sinais de áudio e classificando-o como pertencente a alguma emoção conhecida (IRIYA, 2014, p. 34).

O problema de reconhecimento (de qualquer coisa), pode se dar de duas maneiras: reconhecimento ou verificação. Em ambas as abordagens o reconhecimento de emoções, por exemplo, está limitado pelo universo definido. Quanto maior o universo, maior a complexidade do sistema, e menor a acurácia.

Nilofer et al. (2015) afirma que o reconhecimento de emoções compreende os seguintes passos: entrada de áudio, pré processamento, extração de características, classificação e saída (emoção).

O modelo dimensional do espaço de emoções pode ser utilizado para auxiliar o reconhecimento automático. Logo, “o problema geral de reconhecimento de um conjunto grande de emoções pode ser dividido em problemas menores, se agruparmos as emoções

conforme suas similaridades em cada dimensão" (IRIYA, 2014). Então, pode-se ter um classificador para cada dimensão do espaço emocional, que pode ser executado sequencialmente, iniciando com o classificador da dimensão que conseguir obter a maior taxa de reconhecimento, diminuindo a complexidade do problema para cada uma das fases. Características de voz simples, como volume e frequência fundamental são de extrema importância para a dimensão da ativação, portanto esta é a dimensão mais fácil de ser reconhecida e deve ser a primeira a ser reconhecida, seguida da Avaliação e por último o Domínio.

No sistema de classificação em três estágios descrito por Iriya (2014), o sistema funciona em duas etapas distintas: o treinamento e a classificação. Tanto no treinamento quanto na classificação, uma amostra de sinal de áudio passa primeiramente por um detector de atividade vocal. Esta fragmentação do sinal é realizada para evitar tanto o uso de características de voz incorretos quanto o favorecimento de sinais com longos períodos de silêncio ou ruído. Destes trechos, são extraídas as características de voz relevantes para o processo e estes são usados para treinar o modelo de emoções. Um processo muito semelhante é realizado na fase de classificação, porém após as características de voz serem extraídas, o conjunto de características é confrontado com cada um dos modelos existentes e será classificado pelo modelo que apresenta a melhor coerência.

2.3 Características de Voz

As emoções devem ser modeladas e reconhecidas pelas alterações fisiológicas da voz humana, quando um indivíduo vivencia um determinado estado emocional. Estas alterações podem envolver tanto a prosódia - o estudo do ritmo, entonação e demais atributos correlatos na fala - quanto a qualidade da voz, relacionada à inteligibilidade e à naturalidade da fala.

Apesar de não haver uma regra de quais características utilizar para o reconhecimento, vários autores utilizam algumas em comum. As principais características prosódicas utilizadas são a frequência fundamental (*pitch*), a energia e propriedades relacionadas à duração da voz e pausas, enquanto que entre as características que descrevem a qualidade da voz encontram-se as frequências formantes, a distribuição espectral, representada por MFCC's (*Mel Frequency Cepstral Coefficients*) ou LFPC's (*Log Frequency Power*

Coefficients), a relação harmônicos/ruído (*Harmonic to Noise Ratio* ou HNR) e o fluxo glotal.

2.3.1 Características de Voz Populares

Nesta seção, algumas das características mais populares no estado da arte em reconhecimento de emoções serão brevemente descritos. Isto não indica, porém, que somente as características aqui descritas podem ajudar a obter uma boa acurácia, pois, outras não tão populares podem obter uma precisão bastante relevante, como coeficientes cepstrais delta-espectrais baseados em Teager (T-DSCC) utilizado por Chapaneri e Jayaswal (2013) e coeficientes cepstral preditivos lineares (LPCC) citado por Chandrasekar, Chapaneri e Jayaswal (2014).

2.3.1.1 *Pitch*

O *pitch* pode ser entendido como a altura (frequência) percebida num sinal de áudio. Na maioria das vezes ele se iguala a frequência fundamental conhecida por f_0, porém estes dois podem ser diferentes. Para entendê-lo podemos pensar nas frequências emitidas ao cantarmos uma música. Embora existam várias, é o *pitch* que forma a melodia, logo, é intuitivo pensar que o mesmo pode contribuir para a expressão de emoções. O estudo do *pitch* vem desde a psicologia, e está normalmente associado à dimensão da ativação, sendo que um *pitch* alto com muitas variações reflete emoções de alta excitação.

2.3.1.2 Energia

A energia é a intensidade sonora percebida pelo ouvido humano. O ouvido consegue distinguir a intensidade sonora em decibéis, entretanto é difícil medir a intensidade sonora diretamente. Por este motivo, o que é usado normalmente é a energia do sinal, capturada através da amplitude das amostras. A curva de energia depende de muitos fatores como os fonemas, o locutor, a cultura do locutor, como também do estado emocional do locutor. Como o *pitch*, altos valores de energia estão normalmente correlacionados com emoções de alta excitação.

2.3.1.3 Formantes

As formantes representam picos na resposta em frequência do aparelho fonador humano, isto é, podem ser entendidas como frequências de ressonância do trato vocal humano. Boa parte da característica das vogais é decorrente das duas primeiras formantes, o que as tornam extremamente importantes para o estudo da fonética e da articulação. Como a articulação está bastante correlacionada com o estado emocional, é de se esperar que as formantes tenham influência no reconhecimento de emoções.

2.3.1.4 Coeficientes de Potência em Escala Logarítmica

Os coeficientes de potência em escala logarítmica (LFPC) são coeficientes que refletem a distribuição espectral de energia do sinal. O cálculo destes coeficientes é feito passando-se o espectro do sinal por um banco de filtros que delimitam sub-bandas e calcular a média da energia nestas sub-bandas. Os centros das sub-bandas podem ter distribuição logarítmica para adequação ao modelo psicoacústico da audição humana. LFPCs têm sido utilizados em reconhecimento de emoções, pois se adequam a conclusão de Oudeyer de que a energia em baixas frequências possuem papel importante nas emoções.

2.3.1.5 Coeficientes Mel Cepstrais

Os coeficientes mel cepstrais (MFCC) são uma representação paramétrica do espectro de frequências do sinal de voz. A ideia dos coeficentes mel cepstrais vem do inverso do espectro do sinal. Como o espectro do sinal é obtido através de uma transformada de Fourier do sinal no domínio do tempo, o cepstro é obtido através da transformada inversa de fourier do log espectro. A diferença entre o mel-cepstro e o cepstro é que para os coeficientes mel-cepstrais são aplicados filtros com bandas de frequências igualmente espaçadas na escala mel. O uso da escala mel advém de estudos da Psicoacústica, pois a escala mel se aproxima melhor do sistema auditivo humano, que possui um comportamento não-linear na frequência.

2.3.2 Seleção de Características

Algumas características podem ter maior importância que outras, ou ainda podem ser consideradas redundantes. Considerando isto, a seleção de características mais importantes diminui a complexidade computacional. Na prática, o excesso delas também

pode ser prejudicial no desempenho do sistema pois o conjunto de dados de entrada é finito e pode apresentar tendência.

2.3.3 Técnica de Extração de Características

Conforma afirmam Mierswa e Morik (2005),

> [...] os dados de áudio são séries temporais, onde o eixo y é a amplitude atual medida e o eixo x corresponde ao tempo. Eles são univariados, finitos e equidistantes. Cada elemento x_i da série é composto por dois componentes. O primeiro é o componente de índice, que indica uma posição em linha reta (o tempo, por exemplo). O segundo componente é um vetor m-dimensional de valores que é um elemento do espaço de valores.

Segundo Iriya (2014, p. 42),

> [...] o sinal de áudio $s(n)$ é um sinal digital amostrado com uma determinada frequência de amostragem f_a contendo a voz de um dado locutor. Cada valor $s(n_1)$, $s(n_2)$, $s(n_3)$... etc, representa uma amostra do sinal digital. Este sinal passa por um processo de segmentação, que resulta em diversas janelas $x_j(n)|n = 1, 2, 3...N$, sendo a duração da janela da ordem de dezenas de milisegundos. N denota o comprimento da janela em amostras e j a j-ésima janela

Com ambas definições formais descritas acima, podemos descrever o processo de uma maneira mais informal: Para extrair-se as características, o sinal de áudio é dividido em janelas de M milissegundos sendo que para cada próxima janela é deslocado D milissegundos e gerada a segunda janela. É realizado este processo até que o fim do áudio seja encontrado. As características da voz são extraídas então de cada uma destas janelas separadas anteriormente, gerando um valor ou um conjunto de valores para cada uma das janelas, de acordo com a característica que se deseja extrair.

Esse método de extração de características é um procedimento comum, conforme afirmam Chandrasekar, Chapaneri e Jayaswal (2014), visto que as propriedades estatísticas da forma de onda de fala são observadas como sendo relativamente constantes.

A palavra característica no contexto deste trabalho, seria a abstração de *pitch*, energia, formantes, etc, embora o termo mais correto para estes dados numéricos extraídos que serão utilizados como entrada para os algoritmos de classificação seria parâmetros, ou apenas valores de entrada. Portanto, a partir daqui, os valores extraídos das características serão chamados de parâmetros.

Os parâmetros dinâmicos são apenas o próprio conjunto dos parâmetros de curto prazo (extraído de cada uma das janelas), e ao unirmos todos eles temos uma matriz de parâmetros $P_u(j)$ para cada instância de áudio u. A entrada final para os algoritmos de treinamento e de classificação, $P(j)$, pode ser obtida concatenando-se a matrizes de parâmetros de cada instância de áudio do modelo (IRIYA, 2014):

$$P(j) = [P_1(j), P_2(j), ..., P_U(j)]$$

Já os parâmetros globais, ou parâmetros de médio prazo, são geralmente calculados a partir de propriedades estatísticas dos conjuntos de cada parâmetro, como média, desvio padrão, mediana, máximo, mínimo ou extensão de um determinado contorno de um parâmetro.

2.3.4 Comparação entre Características

Vários autores já estudaram as características que podem ser extraídas da voz, porém não há um consenso entre os estudos. Isto se deve à algumas variáveis, como base de dados, conjuntos de emoções reconhecidas e métodos de classificação.

Nwe, Foo e Silva (2003 apud IRIYA, 2014), comparam o desempenho de LPCC, MFCC e LFPC para o reconhecimento de seis emoções utilizando HMMs. Os autores apontam taxas de reconhecimento de 77%, 56% e 51% para LFPCs, MFCCs e LPCC's respectivamente.

O *pitch*, a priori, é considerado uma característica extremamente importante para o reconhecimento, porém (LIN; WEI, 2005 apud IRIYA, 2014) afirmam que os valores do *pitch* em si são menos relevantes do que suas alterações instantâneas, como por exemplo suas derivadas e sua energia.

2.4 Métodos de Classificação

Como já dito anteriormente, o reconhecimento de emoções através da voz tem sido resolvido como um problema estatístico ou de reconhecimento de padrões, onde um modelo é gerado extraindo-se características de voz dos áudios, cujas emoções são conhecidas e posteriormente uma instância cuja emoção é desconhecida é confrontada com os modelos existentes e o modelo mais adequado é escolhido. Os parâmetros de voz

podem ser globais ou dinâmicos. Estes dois tipos deram origem à dois tipos de classificação: estática, que utiliza parâmetros de médio prazo e dinâmica que utiliza o conjunto dos parâmetros de curto prazo.

Os parâmetros globais são necessários, por conta de que parâmetros de curto prazo não representam individualmente uma informação útil para o modelo em si. Por outro lado, uma desvantagem da classificação estática é que como são geradas estatísticas globais do contorno, a informação temporal é perdida.

Já na classificação dinâmica utiliza-se todo o contorno dos parâmetros de curto prazo tanto para treinar os modelos quanto para a classificação em si. A classificação dinâmica é mais comum para métodos estatísticos, uma vez que toda a sequência de parâmetros de curto prazo pode ser utilizada para alimentar o algoritmo (IRIYA, 2014).

2.4.1 Comparação Entre Métodos de Classificação

Ambos os métodos de classificação podem ser eficientes para diferentes tipos de problemas, e só seria possível compará-los em um mesmo tipo de estudo sendo, portanto, difícil encontrar dois estudos tão similares utilizando métodos diferentes. Há autores que desenvolveram também sistemas híbridos que alcançaram desempenhos ligeiramente superiores.

2.4.2 Algoritmos de Classificação

Neste tópico serão apresentados apenas os métodos de classificação mais populares e que alguns autores já confirmaram que apresentam bons resultados para o tipo de classificação que pretendemos implementar neste trabalho, como o trabalho realizado por Le e Provost (2013) utilizando HMM e o sistema para reconhecimento de mandarim realizado por Pao, Chen e Yeh (2006) que alcançou incríveis 84.2% de precisão.

2.4.2.1 Modelos de Misturas de Gaussianas

Modelos de Misturas de Gaussianas (GMM) são um tipo particular de modelos de misturas, cuja importância é indiscutível para a área de processamento de voz e principalmente para os temas de reconhecimento de voz e de locutor.

GMM's são largamente utilizados para estimar funções densidade de probabilidade desconhecidas, onde a tarefa de estimação torna-se a de estimar os pesos de cada gaussiana e suas respectivas médias e matrizes de covariância a partir dos dados observados. Podem também ser utilizados como mecanismo de classificação bayesiana, onde é escolhido o modelo que apresenta maior verossimilhança em relação à sequência de observações.

Para GMM são preferíveis a utilização de parâmetros de curto prazo, porém há autores que usaram parâmetros globais (médio prazo).

2.4.2.2 Modelos Ocultos de Markov

Modelos Ocultos de Markov (HMM) apresentam um grande potencial para modelar as emoções humanas, uma vez que possuem o poder de manter informações temporais da sequência de observações através do uso de estados.

HMM's são modelos estatísticos, associados ao Processo de Markov e às Cadeias de Markov. Neles, a ocorrência de uma sequência de eventos observáveis é decorrente de os eventos terem percorrido uma sequência de estados, sendo que cada estado emite um símbolo ou conjunto de símbolos observáveis, segundo alguma distribuição probabilística.

O Processo de Markov é um processo estocástico controlado por uma variável aleatória que representa o estado do sistema, cujo valor futuro depende dos estados passados. No processo de Markov de primeira ordem, o valor futuro depende apenas do estado atual, não da sequência de estados que o procederam, propriedade conhecida como Propriedade de Markov.

2.4.2.3 K-Vizinhos Mais Próximos

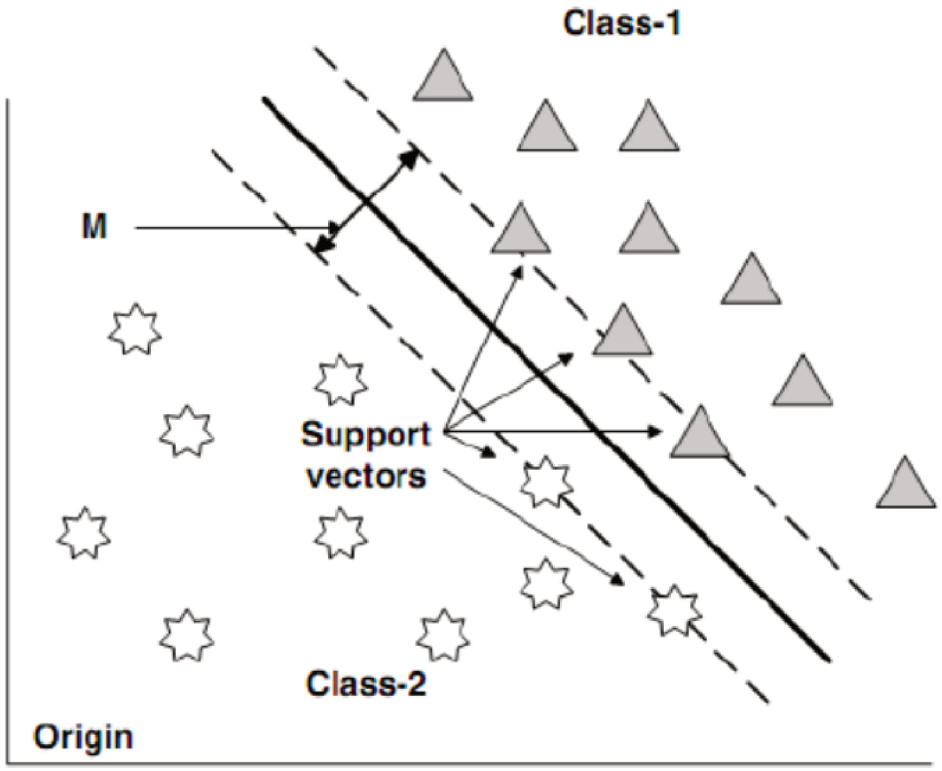
O K-Vizinhos Mais Próximos (KNN) é um algoritmo não-paramétrico utilizado em reconhecimento de padrões com parâmetros globais. É um dos algoritmos mais simples utilizados em aprendizagem de máquina, no qual não são gerados modelos propriamente ditos. Em vez disso, a fase de treinamento consiste simplesmente em armazenar o vetor de parâmetros das amostras de treinamento e suas conhecidas classes. Uma amostra de teste cuja classe é desconhecida, é então classificada de acordo com sua proximidade às amostras de treinamento, que são chamadas de vizinhos. A classe em que a amostra será classificada é aquela que se repete mais vezes dentre os vizinhos mais próximos.

Entretanto, o KNN é um algoritmo muito simples e está longe de ser a melhor solução para um problema de reconhecimento de padrões.

2.4.2.4 Máquinas de Vetores de Suporte

Máquinas de Vetores de Suporte (SVM) é uma técnica utilizada principalmente para reconhecimento de padrões, onde segundo Chavhan, Dhore e Yesaware (2010), normalmente é utilizado como classificador binário, porém também pode ser utilizado como um classificador multiclasses. As SVMs têm sido largamente utilizada para o reconhecimento de emoções através da voz. Elas fornecem uma solução ótima para o problema de separação de duas classes linearmente separáveis, então para o reconhecimento de apenas fazer parte ou não da classe de uma emoção provavelmente apresentaria bom desempenho.

Figura 3 Separação de duas classes com SVM.



Fonte: Jarande e Waghmare (2015)

Segundo Jarande e Waghmare (2015), o intuito básico de SVM é criar um hiperplano, onde seu objetivo é separar as entradas em duas classes. A margem M é a distância entre os dois pontos mais próximos das duas diferentes classes. O classificador SVM posiciona a borda de decisão utilizando a margem máxima entre todos os possíveis hiperplanos.

3 Trabalhos Relacionados

Nesta seção são apresentados de forma sucinta o desenvolvimento de alguns trabalhos principais que estão diretamente relacionados ao sistema que será proposto. O foco destes trabalhos também é o reconhecimento de emoções, sendo que um deles utiliza o mesmo banco de dados que será utilizado no sistema proposto para o presente trabalho de conclusão de curso.

3.1 *A Comprehensive Survey on Features and Methods for Speech Emotion Detection*

Alva, Nachamai e Paulose (2015) descrevem que computação emocional é uma área da inteligência artificial que busca preencher o gap entre emoções humanas e tecnologia da computação. Este campo é multidisciplinar, envolvendo estudos das áreas de ciência da computação, ciência cognitiva e psicologia. As emoções guiam a maneira como é percebido o mundo ao redor do indivíduo, logo uma máquina que consegue compreender as emoções humanas poderá dar uma resposta mais adequada à emoção identificada.

Características prosódicas são aquelas que tratam do aspecto musical da fala, como por exemplo o ritmo e entonação da fala.

No experimento conduzido por Bjorn Schiller analisado pelos autores deste trabalho, foi utilizado o *pitch* e a energia. Foram realizados três diferentes testes com o mesmo conjunto de dados: o primeiro com CHMM, que alcançou uma taxa de reconhecimento de 77,8%; o segundo utilizando *Gaussian Mixture Model* (GMM), que obteve 86,8% de taxa de reconhecimento; e o terceiro que foi a análise por humanos, que alcançou uma média de taxa de reconhecimento de 81,3%.

A taxa de classificação de emoções utilizando características prosódicas estão praticamente iguais ao nível de classificação por humanos, porém o problema do uso dessas características é que elas são dependentes de linguagem e não podem ser usado de forma universal (ALVA; NACHAMAI; PAULOSE, 2015).

Características que lidam com o som envolvendo a fala são fonéticas. Kishore e

Satish (2013 apud ALVA; NACHAMAI; PAULOSE, 2015) sugeriu utilizar características baseadas na fonética tais como MFCC e características *wavelet* para reconhecimento de emoções através da fala. Foram consideradas 6 emoções: raiva, nojo, medo, alegria, neutro e tristeza. Os resultados mostraram que *Sub-Band Cepstrum* (SBC) superou com 70% de precisão contra 51% obtidos em MFCC, pelo fato de SBC ser menos sensível ao ruído.

Chee et al. (2009 apud ALVA; NACHAMAI; PAULOSE, 2015) conseguiram obter precisão próxima aos 90% utilizando KNN e Análise Discriminante Linear (em inglês, *Linear Discriminant Analysis* ou LDA). T L New et al. utilizou em seu estudo LFPC e outras características baseadas no domínio do tempo e frequência, como NFD-LFPC, NTD-LFPC, entre os quais LFPC foi a que obteve melhor performance (87,8%). Uma variação do HMM chamado HMM Contínuo (CHMM) foi utilizado para classificação.

Alva, Nachamai e Paulose (2015) afirmam que características como pitch, duração, MFCC e TEO são utilizados para detectar emoções transmitidas através da fala. E também, que classificadores como HMM e ANN têm sido extensivamente estudados pela sua eficiência.

3.2 *Emotion Recognition From Spontaneous Speech Using Hidden Markov Models With Deep Belief Networks*

Uma rede de aprendizado profunda (*Deep Belief Network* ou DBN) consiste em uma pilha de Máquinas Restritas de Boltzmann (RBMs) treinadas gradualmente camada por camada (LE; PROVOST, 2013). RBMs contém dois conjuntos de unidades visíveis e ocultas interconectadas, e podem ser de dois tipos: Bernoulli e Gaussiana. Nas RBMs Bernoulli, tanto as camadas ocultas como as visíveis são binário. Nas RBMs Gaussianas, as unidades visíveis podem ser um número real.

As camadas ocultas de uma RBM pode ser a entrada de outra RBM que é treinada separadamente do modelo anterior. Esta pilha de RBMs generativas pré-treinadas constituem uma DBN, que pode ser ajustada como uma Rede Neural Artificial (RNA).

Le e Provost (2013) propõem uma modelagem dinâmica a nível do *frame*, usando modelos acústicos baseados em DBN em conjunto com HMMs. Foi utilizado *Hidden Markov Toolkit* (HTK) para extração das características MFCC de cada uma das

expressões utilizando uma janela de 25ms e um passo de 10ms. Foi realizado também uma z-normalização sobre cada locutor do banco de dados FAU Aibo, que foi o BD utilizado pelo autor.

Os autores, em seu estudo, treinaram um conjunto de classificadores híbridos utilizando HHMs para capturar propriedades temporais das emoções DBNs para estimar as probabilidades.

Os modelos avaliados pelo autor foram três HMM (com 1, 3 e 5 estados) e 9 tamanhos de janelas de contexto. Quando combinados, resultaram, na verdade em 27 modelos avaliados. "Não há melhor arquitetura para todas as emoções e expressões" (LE; PROVOST, 2013). Isto sugere que combinar arquiteturas é uma boa estratégia, pois cada arquitetura pode ser eficaz para reconhecer uma emoção diferente.

Foi produzido uma *unweighted average recall* (UAR) máximo de 46,36%, que são valores maiores do que alguns estudos existentes na literatura.

3.3 Análise de Sinais de Voz para Reconhecimento de Emoções

O estudo conduzido por Iriya (2014) utilizou uma base de dados atuada chamada *Berlin Emotional Database of Speech* (EMO-DB), que é composta de sete emoções: Felicidade, Medo, Neutro, Nojo, Raiva, Tédio e Tristeza.

O autor selecionou para seu estudo um conjunto de 96 características de curto prazo, incluindo *pitch*, energia, as 5 primeiras formantes, os 13 primeiros MFCC e os 12 primeiros LFPC, além das primeiras e segundas derivadas de cada características.

Os parâmetros são extraídos em janelas de 40 ms, devido à necessidade de a janela ser duas a três vezes maior que o período de pitch máximo para extração do pitch (BOERSMA, 1993 apud IRIYA, 2014). As janelas possuem sobreposição de 75%, sendo calculadas em passos de 10 ms.

Para a extração do *pitch* foi utilizado o algoritmo desenvolvido por Boersma (1993 apud IRIYA, 2014). Foi escolhido por ser um algoritmo relativamente simples, com boa eficiência e ser extensivamente utilizado por pesquisadores de diversas áreas. É um algoritmo baseado em autocorrelação, que na verdade estima o período de *pitch*, obtendo o número de amostras entre os máximos de correlação do sinal, ou seja, o atraso T.

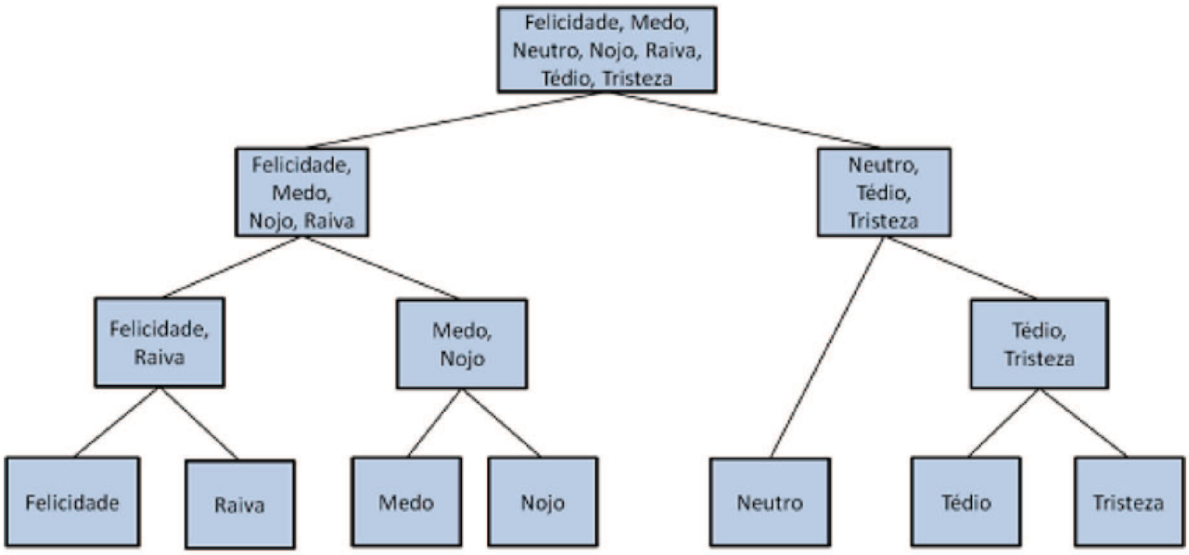
Para a extração da energia de curto prazo, é calculado o log da média da potência do sinal a cada janela.

Iriya (2014) utilizou GMMs, HMM, KNN e SVM em seus testes, sendo a maior parte do trabalho desenvolvida utilizando GMMs. Foi empregada a seleção de características para GMMs que utilizam parâmetros dinâmicos, isto é, o contorno dos parâmetros de curto prazo. É utilizado um algoritmo iterativo chamado *Expectation Maximization* que é comumente usado para problemas de maximização de verossimilhança. É composto de duas etapas, a etapa de Esperança, na qual é calculada a verossimilhança e a etapa de Maximização, na qual os parâmetros das componentes gaussianas são estimados e o modelo é modificado de modo a maximizar a verossimilhança.

HMM e GMM obtiveram os melhores desempenhos gerais entre os métodos testados. Porém, um melhor desempenho geral não necessariamente acompanha um melhor desempenho individual de cada emoção (IRIYA, 2014).

Foi utilizado o modelo dimensional do espaço de emoções para auxiliar o reconhecimento automático, pois desta forma tornou-se possível quebrar o problema de reconhecimento de um conjunto grande de emoções em problemas menores, agrupando as emoções similares no espaço dimensional (Ativação, Avaliação e Domínio). Na figura 4 é apresentado o esquema de classificação em três estágios.

Figura 4 Esquema de classificação em três estágios, proposto por Iriya (2014).



Fonte: Iriya (2014)

É importante destacar que as características selecionadas têm uma taxa de acerto diferente para homens e mulheres, onde utilizando somente *pitch* e energia, para homens foi possível alcançar uma taxa de acerto de 59,23% e para mulheres de 61,26%. Nos testes com o classificador de único estágio, os coeficientes LFPC de baixa ordem foram escolhidos para o caso das locutoras mulheres, em detrimento da energia e a primeira derivada do *pitch* para os locutores homens. Os resultados deste estágio (de 74,86% masculino e 71,82% feminino) foram bastante superiores aos dos primeiros testes. Já na classificação sequencial de três estágios, as taxas de reconhecimento foram ainda melhores, alcançando os 82,41% para homens e 81,28% para mulheres.

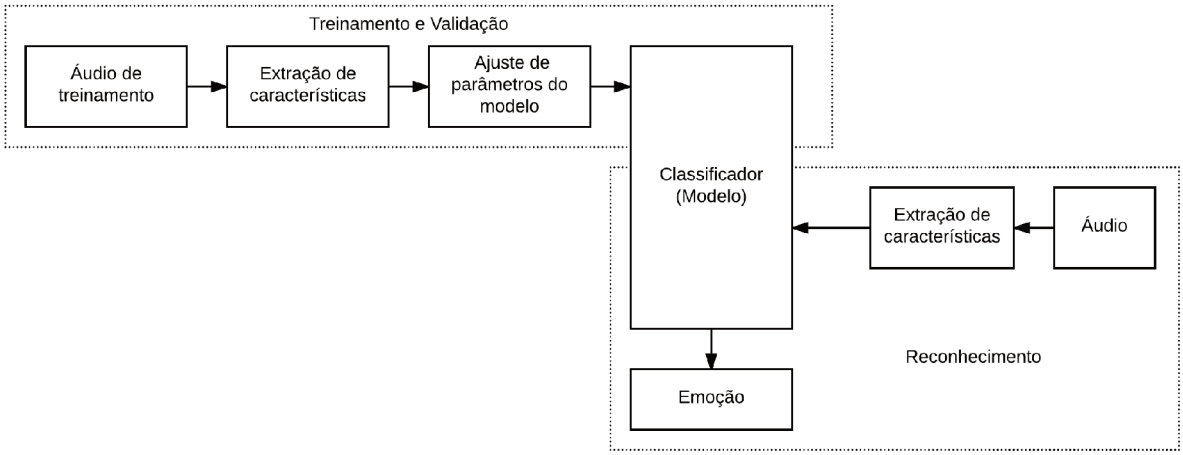
Quanto à seleção de características, é relevante destacar que as formantes aparecem em algumas emoções, o que não é comum no estado da arte. Em contraponto o *pitch* influencia bastante o reconhecimento da emoção felicidade, o que confirma a visão tradicional desta característica.

4 Proposta

O sistema proposto contempla 2 módulos: módulo de entrada (extração de características e geração e validação do modelo) e um módulo de classificação (identificação da emoção).

O sistema de reconhecimento automático de emoções proposto pode ser construído de diversas formas, porém a intenção é desenvolvê-lo de forma que obtenha a maior precisão possível. Para isto se tornar possível é necessário entender como funciona cada um dos métodos de classificação, quais as técnicas utilizadas, e quanto o conjunto de emoções influencia na taxa de reconhecimento, bem como quais os métodos de classificação já existem implementados, pois não é escopo deste trabalho implementar algoritmos de classificação ou extração de características. Um esquema de como se pretende desenvolver este sistema para reconhecimento automático de emoções é apresentado na figura 5.

Figura 5 Principais etapas do sistema proposto.



Fonte: Autoria própria.

4.1 Emoções

Vários autores utilizam conjuntos de emoções ligeiramente diferentes. No estudo realizado por Ververidis e Kotropoulos (2006) as emoções mais comuns que aparecem nos banco de dados pesquisados são: Raiva (*Anger*), Alegria (*Hapiness*, *Joy*), tristeza (*Sadness*), Medo (*Fear*), Surpresa (*Surprise*) e estresse (*Stress*). Porém, alguns destes

bancos de dados nem mesmo foram criados com o propósito de reconhecimento de emoções, e podem não ser adequados para este uso.

No presente trabalho, será desenvolvido um sistema para reconhecer um conjunto limitado de emoções, visando aumentar a precisão geral do reconhecimento. Pois, segundo Iriya (2014), deve-se definir um conjunto mais limitado possível somente com as emoções mais básicas, que sejam de fácil identificação do ponto de vista humano, e principalmente àquelas cujo reconhecimento automático tenha aplicação prática. Portanto, serão utilizadas apenas 4 emoções: Neutro, Raiva, Felicidade e Tristeza.

4.2 Banco de Dados de Áudio

Ververidis e Kotropoulos (2006) realizaram um levantamento dos banco de dados disponíveis que podem ser utilizados para a tarefa de reconhecimento de emoções, porém não foi encontrado nenhum disponível na língua portuguesa.

Tendo em vista que um dos objetivos específicos deste trabalho, descrito na seção 1.2.1, é desenvolver a aplicação para a língua portuguesa e não foi obtido êxito na procura de bancos de dados adequados para este fim nesta linguagem, optou-se por extrair trechos de áudios de filmes e vídeos brasileiros manualmente, visando cumprir este objetivo. Mesmo com a base em Português, foi utilizado também o banco de dados de emoções de Berlin (EMO-DB) para realizar os testes iniciais, pois é um banco de dados bastante estável, utilizado em diversos estudos como o de Iriya (2014), e o de Alva, Nachamai e Paulose (2015).

4.2.1 Banco de Dados Áudios de Emoções de Berlin

O banco de dados EMO-DB[1] é atuado, ou seja, não foi capturado em uma situação em que a emoção realmente aconteceu, mas foi gravado por atores em um estúdio. Ele tem aproximadamente 530 áudios, com 10 frases diferentes faladas em alemão, sendo que estas frases não possuem nenhum significado que influencie no reconhecimento da emoção. Estas frases foram faladas por 5 locutoras mulheres e 5 locutores homens.

Conforme afirma Nilofer et al. (2015), à medida que a naturalidade do banco de dados aumenta, a complexidade também aumenta. As emoções do EMO-DB foram

[1] http://emodb.bilderbar.info/docu/.

relativamente fáceis de se obter por ser um banco de dados atuado, o que intuitivamente diminui a margem para erro da classificação realizada pelos criadores deste banco de dados. Os tipos de bancos de dados e seu nível de dificuldade para criação podem ser observados na figura 6.

Figura 6 Tipos de banco de dados para reconhecimento de emoções e seu nível de dificuldade.

Fonte: Nilofer et al. (2015)

No banco de dados são considerados o estado neutro (63 amostras) mais 6 emoções, sendo elas raiva (101 amostras), tédio (65 amostras), nojo (37 amostras), ansiedade/medo (55 amostras), felicidade (57 amostras) e tristeza (50 amostras). Para os nossos testes, foram utilizados apenas as emoções raiva, felicidade, tristeza e neutro deste banco de dados pois correspondem às emoções dos áudios obtidas em português. Também balanceamos a quantidade de amostras que existem para cada uma das emoções, deixando cada emoção com a mesma quantidade de amostras de áudio.

4.2.2 Banco de Dados em Português

A criação do banco de dados na língua portuguesa[2] foi realizada extraindo-se trechos de filmes ou vídeos do YouTube[3] com apenas uma frase e poucos segundos de duração. Todos os áudios foram extraídos utilizando o software Audacity[4].

A classificação foi realizada de forma manual pelo autor deste trabalho, através da análise dos trechos de áudio extraídos e escolha da emoção que apresentava a maior similaridade com as emoções do escopo deste trabalho. Devido à separação em classes destes áudios ter sido feita de forma manual, poderá haver uma taxa de erro associado à esta classificação.

[2] https://github.com/jdarosaj/emotion_portuguese_database
[3] https://www.youtube.com/
[4] http://www.audacityteam.org/home/

O banco de dados contém 37 áudios que representam a emoção raiva, 27 a emoção felicidade, 22 a emoção tristeza e 24 áudios que não expressam emoções, sendo considerados neutros, completando 110 amostras no total.

4.3 Características de Voz

As características utilizadas na grande maioria dos estudos encontrados são o *pitch* e a energia, pois é sabido que ambos podem carregar bastante informação (SCHULLER; RIGOLL; LANG, 2003). Com isto, pode-se esperar que sejam adequados para o reconhecimento de emoções.

Segundo Iriya (2014, p. 19),

> [...] vários autores constatam que as características da voz têm papéis diferentes para o reconhecimento de emoções em locutores do sexo masculino e feminino, o que faz sentido se pensarmos que a frequência fundamental das mulheres é geralmente mais alta do que a dos homens.

Não é necessário, apenas, determinada característica carregar bastante informação para ser útil no reconhecimento, ela deve ser adequada ao gênero do locutor. Porém, nem sempre é possível saber de antemão se a voz analisada é masculina ou feminina, para isto seria necessário realizar uma identificação de gênero antes de realizar o reconhecimento da emoção. Não está no escopo deste trabalho realizar identificação do gênero do interlocutor.

4.3.1 Extração de Características

A extração das características previamente escolhidas é realizada em cada uma das janelas de áudio que são separadas para análise. Inicialmente utilizaremos o tamanho da janela igual a 36 ms, por estar na faixa de 20 à 100 ms determinada a mais adequada por Giannakopoulos (2015).

Segundo Schuller, Rigoll e Lang (2003), os valores da energia são calculados pela média logarítmica da energia dentro de uma janela. E o contorno do *pitch* é obtido através do uso da *average magnitude difference function* (AMDF, em português chamada de função de diferença da magnitude média).

O sistema desenvolvido neste trabalho, faz uso da biblioteca *pyAudioAnalisys*[5] criada por Giannakopoulos (2015), que utilizou outra biblioteca desenvolvida por Pedregosa et al. (2011) chamada *scikit-learn*, para o treinamento de SVM. A *pyAudioAnalisys* realiza a extração das características do áudio apresentados na tabela 1.

Tabela 1 – Características extraídas pela biblioteca *pyAudioAnalisys*.

NOME DA CARACTERÍSTICA	**DESCRIÇÃO**
Taxa *Zero Crossing*	A taxa de mudança do sinal durante a duração de um frame em específico.
Energia	A soma dos quadrados dos valores do sinal, normalizados pelo respectivo comprimento do *frame*.
Entropia da Energia	A entropia das energias normalizadas dos sub-frames. Pode ser interpretado como uma medida de mudanças rápidas.
Centróide Espectral	O centro de gravidade do espectro.
Espalhamento do Espectro	O segundo momento central do espectro.
Entropia Espectral	Entropia das energias espectrais normalizadas para um conjunto de *sub-frames*.
Fluxo Espectral	A diferença quadrática entre as magnitudes normalizadas do espectro de dois *frames* sucessivos.
Deslocamento espectral	A frequência abaixo da qual se concentra 90% da distribuição de magnitude do espectro.
MFCC	Os coeficientes cepstrais de frequência Mel formam uma representação cepstral onde as bandas de frequência não são lineares, mas distribuídas de acordo com a escala Mel.
Vetor de *Chroma*	Uma representação da energia espectral com 12 elementos em que os recipientes representam 12 classes de *pitch* com comportamento constante de músicas ocidentais (espaçamento de semitom).
Desvio de *Chroma*	O desvio padrão dos 12 coeficientes de *chroma*.

Fonte: Giannakopoulos (2015)

4.4 Classificadores

Optou-se pela utilização de KNN e SVM para os testes iniciais, por serem classificadores simples, amplamente implementados, fáceis de encontrar, e que não necessitam de muita capacidade de processamento, se comparados à HMM e GMM. Esta decisão também foi influenciada pela escolha de outros autores, como Iriya (2014), já validaram a

[5] https://github.com/tyiannak/pyAudioAnalysis

eficácia de HMMs e GMMs, mas principalmente por influência do trabalho de Sujatha e Ameena (2016) onde é mostrado que, em reconhecimento de emoções dependente do texto, o SVM é 15% a 20% superior aos classificadores citados anteriormente, verificaremos se o reconhecimento independente de texto também apresenta um bom resultado.

O classificador SVM é, essencialmente, utilizado para classificação binária, mas existem implementações para classificação multiclasse. Na próxima seção serão sucintamente abordadas as duas principais técnicas utilizadas na literatura.

4.4.1 Técnicas de Classificação

A biblioteca *scikit-learn*[6] utiliza uma técnica de classificação chamada *"one-against-one"* para classificação não-binária, onde considerando N como sendo o número de classes a ser reconhecida, então são gerados N * (N - 1) / 2 classificadores, cada um deles treinado com amostras de duas classes.

Existe também outra técnica empregada chamada *"one-against-all"* empregada por alguns autores, como Liu e Zheng (2005), onde argumentam que esta estratégia consiste em construir um classificador SVM por classe, que é treinado para distinguir as amostras de uma classe entre as amostras de todas as outras classes restantes.

Porém, segundo Milgram, Cheriet e Sabourin (2006), esta técnica apresenta uma precisão um pouco maior somente com pouca quantidade de classes a ser reconhecida, enquanto que a estratégia utilizada pela biblioteca *"one-against-one"* apresenta uma acurácia boa também para o reconhecimento de várias classes.

4.5 Escolha do Modelo

Para a escolha do modelo que será utilizado no presente trabalho, utilizaremos um esquema de validação cruzada (do inglês *cross-validation*), onde será estimado qual o melhor valor de parâmetro para o classificador, no caso do SVM, o parâmetro de custo C e no caso do KNN, o número de vizinhos K.

Validação cruzada é uma técnica para avaliar a capacidade de generalização de um modelo, a partir de um conjunto de dados[7]. É bastante utilizada para analisar

[6] http://scikit-learn.org/stable/
[7] https://pt.wikipedia.org/wiki/Validaçao_cruzada

problemas de predição, visando estimar o quão preciso é um modelo de predição.

5 Experimentos Práticos

Neste capítulo serão descritos os resultados e algumas decisões ao longo dos experimentos realizados, bem como os parâmetros utilizados para treinamento e, ao final, será apresentada uma comparação entre as duas estratégias.

5.1 Critérios para Utilização do Modelo

Para cada um dos modelos treinados utilizamos várias configurações de parâmetros, e a cada rodada de treinamento/testes é calculada uma matriz de confusão para os testes de validação. Esta matriz permite a visualização do desempenho do algoritmo, onde cada linha da matriz representa as instâncias de uma classe predita, enquanto que cada coluna representa as instâncias da classe real[8]. Neste trabalho apresentaremos em porcentagem ao invés de valores absolutos, visando facilitar sua interpretação.

Para definirmos qual o parâmetro que resultou em um melhor treinamento, é utilizada a métrica de acurácia, sendo esta a proporção de predições corretas, que é obtida através do seguinte cálculo:

$$(\sum d / \sum e) * 100$$

Onde d é cada elemento da diagonal da matriz de confusão, que representam os acertos. Já o e representa cada um dos elementos da matriz de confusão representando todo o conjunto de testes. O resultado deste cálculo será um número de 0 a 100 representando a porcentagem de acurácia do modelo.

5.2 Testes Iniciais com Diferentes Classificadores

Os primeiros testes foram realizados apenas com o Berlin Emo-DB visto que este banco de dados já foi extensivamente testado e utilizado por diversos autores, sendo este bastante estável. Foram realizados testes com diferentes classificadores a fim de encontrar

[8] https://en.wikipedia.org/wiki/Confusion_matrix

qual obteve a melhor performance. Todos os testes iniciais foram realizados utilizando 75% dos áudios disponíveis (separados aleatoriamente) para treinamento dos modelos e os outros 25% para testar os modelos gerados com os diversos parâmetros.

5.2.1 Testes Utilizando KNN

Apesar se ser um método de classificação bastante simples, o KNN apresentou resultados razoáveis. Os testes com KNN foram realizados com os seguintes valores de parâmetros:

- K: 1, 3, 5, 7, 9, 11, 13, 15;
- Janela de curto prazo: 36 ms;
- Passo da janela de curto prazo: 12ms;
- Janela de médio prazo: 400 ms;
- Passo da janela de médio prazo: 125ms;

O primeiro teste com KNN foi realizado utilizando todas as emoções disponíveis no banco de dados EMO-DB. O parâmetro que obteve a melhor acurácia foi $K = 3$, com 66.7% de acurácia, resultando na matriz de confusão apresentada na tabela 2.

Tabela 2 – Matriz de confusão para KNN com todas as emoções.

	Raiva	**Tédio**	**Nojo**	**Medo**	**Alegria**	**Neutro**	**Tristeza**
Raiva	21,26	0,02	0,44	0,03	1,77	0,06	0,00
Tédio	0,08	10,42	0,25	0,23	0,01	3,26	0,84
Nojo	1,22	0,32	4,91	0,97	0,13	0,72	0,23
Medo	2,18	1,38	2,25	5,81	0,89	0,58	0,11
Alegria	4,47	0,73	0,79	0,86	5,92	0,43	0,00
Neutro	0,09	4,62	0,29	0,16	0,32	8,74	0,87
Tristeza	0,00	1,11	0,26	0,27	0,00	0,05	9,62

O segundo teste foi realizado utilizando somente as emoções que pretendemos utilizar nos áudios em português, que são alegria, tristeza, raiva e neutro. Os parâmetros utilizados foram os mesmos do teste anterior. O melhor parâmetro K obtido neste segundo

Tabela 3 – Matriz de confusão para KNN com quatro emoções.

	Raiva	**Alegria**	**Neutro**	**Tristeza**
Raiva	34,07	2,88	0,36	0,00
Alegria	7,84	11,48	1,58	0,00
Neutro	0,33	1,04	19,81	2,70
Tristeza	0,00	0,00	1,07	16,84

teste também foi $K = 3$, onde foi obtido 82.2% de acurácia. A matriz de confusão que foi obtida é exibida na tabela 3.

A partir desses testes realizados é possível validar o que foi estudado anteriormente. Quanto maior for o universo de classes a ser reconhecido, maior a complexidade do sistema e consequentemente, menor a acurácia. Com isso se confirma a hipótese que tínhamos, de utilizar um conjunto limitado de emoções para não haver muito impacto na acurácia do modelo.

5.2.2 Testes Utilizando SVM

As máquinas de Vetores de Suporte normalmente apresentam bons resultados para decisões binárias, como por exemplo classificar um áudio como uma emoção negativa ou positiva, porém decidimos realizar alguns testes utilizando classificação multiclasse para haver uma comparação entre dois modelos de classificação diferentes.

O primeiro teste com SVM foi realizado utilizando os mesmos valores de janela e passo do teste para KNN, a fim de se fazer uma comparação entre os dois classificadores. Foram testados 0.001, 0.01, 0.5, 1.0, 5.0, 10.0, 20.0 como os valores de custo C, sendo o parâmetro 10.0 o que obteve a melhor acurácia (71.5%). A matriz de confusão para este teste é apresentada na tabela 4.

Na figura 7 é apresentada graficamente uma comparação da taxa de acerto entre os dois classificadores testados.

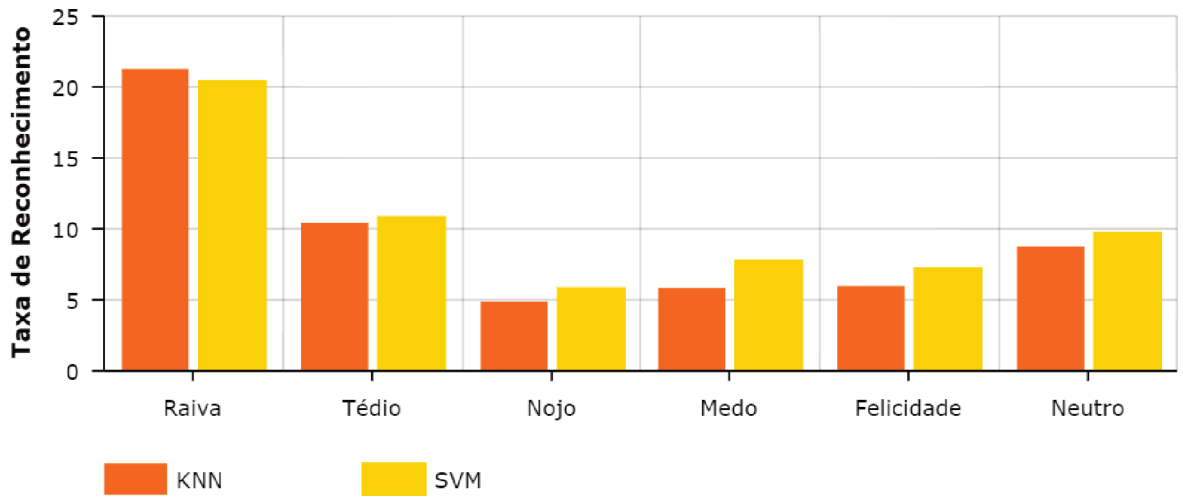
É possível perceber que SVM já apresentou uma melhor precisão do que KNN neste primeiro teste, alcançando 71.4% contra 66.7% do KNN.

No segundo teste realizado, utilizamos os mesmos parâmetros, porém utilizamos os áudios somente das 4 emoções que pretendemos reconhecer. A matriz de confusão obtida

Tabela 4 Matriz de confusão para SVM com todas as emoções.

	Raiva	**Tédio**	**Nojo**	**Medo**	**Alegria**	**Neutro**	**Tristeza**
Raiva	20,37	0,00	0,25	0,24	2,74	0,00	0,00
Tédio	0,00	10,79	0,16	0,55	0,07	3,20	0,33
Nojo	0,62	0,06	5,86	1,27	0,35	0,27	0,06
Medo	0,75	0,54	1,02	7,83	1,64	0,82	0,61
Alegria	3,22	0,23	0,87	1,35	7,20	0,35	0,00
Neutro	0,00	3,55	0,39	0,63	0,21	9,78	0,54
Tristeza	0,00	0,73	0,04	0,26	0,00	0,54	9,75

Figura 7 Comparação da taxa de reconhecimento entre KNN e SVM.



Fonte: Autoria própria.

é exibida na tabela 5.

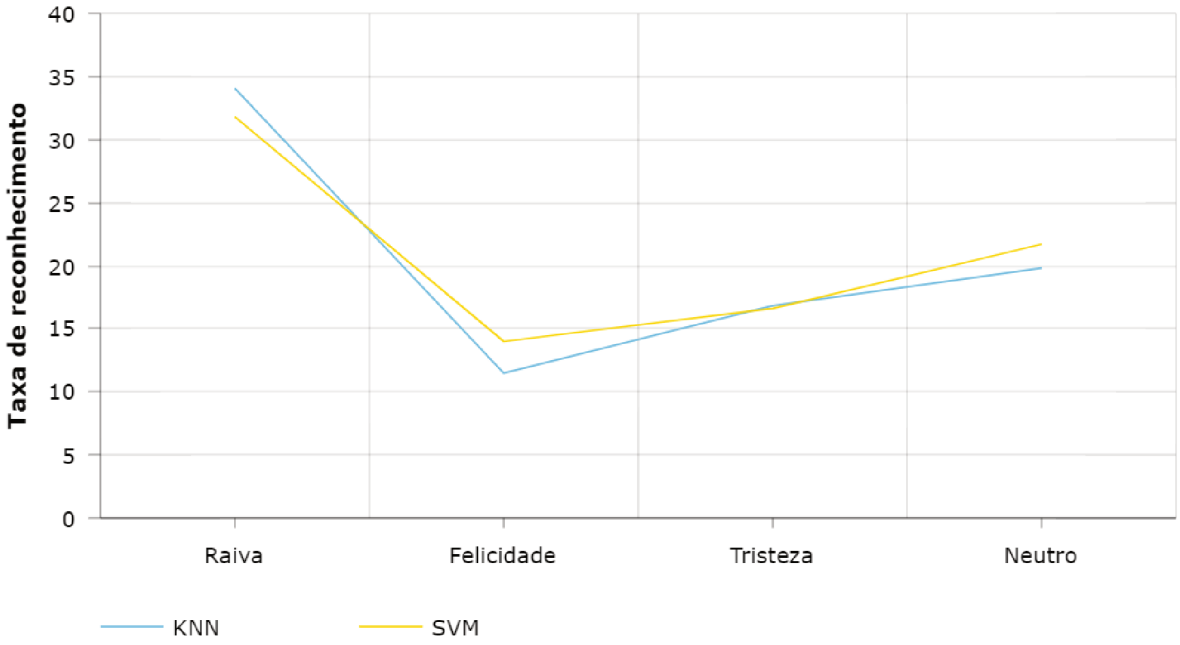
Tabela 5 Matriz de confusão para SVM com quatro emoções.

	Raiva	**Alegria**	**Neutro**	**Tristeza**
Raiva	31,81	5,46	0,04	0,00
Alegria	5,64	14,00	1,24	0,01
Neutro	0,00	0,82	21,70	1,36
Tristeza	0,00	0,00	1,31	16,60

É possível analisar na figura 8, que as emoções felicidade e neutro obtiveram melhor desempenho com o classificador SVM, enquanto que na emoção tristeza o desempenho foi similar, sendo o KNN ligeiramente superior apenas na emoção raiva.

A taxa de reconhecimento deste modelo foi de 84.1%, contra 82.2% no método

Figura 8 Desempenho dos classificadores SVM e KNN para apenas 4 emoções.



Fonte: Autoria própria.

KNN. Podemos perceber que o *gap* entre a precisão de um método e outro diminuiu bastante, apesar de SVM ainda ser mais preciso do que do método anterior.

Nos próximos testes utilizaremos apenas a classificação com SVM, por apresentar melhor precisão geral se comparado ao KNN.

5.3 Problema de Desbalanceamento da Quantidade de Amostras

Após alguns experimentos percebemos que a quantidade de amostras de cada uma das emoções eram muito discrepantes. Para resolver este problema tínhamos duas alternativas:

- Remover alguns áudios das emoções com mais amostras, com o mesmo objetivo anterior;
- Duplicar os áudios das emoções com menos amostras, visando equilibrar o número de áudios entre todas as emoções.

Utilizando a primeira alternativa, perderíamos uma grande quantidade de dados novos que podem agregar bastante informação ao modelo treinado, o que logicamente

afetaria negativamente o modelo gerado.

Porém, ao utilizar a segunda alternativa, não conseguiríamos equilibrar exatamente o número de amostras de áudio, pelo simples fato de não podermos duplicar somente apenas alguns áudios, pois isto levaria à um desbalanceamento do modelo no universo desta emoção. Optamos por duplicar os áudios das emoções com menos amostras, a fim de não perder nenhuma informação que possa ser útil ao modelo, pagando o preço de a quantidade de amostras continuar levemente desbalanceada: 101 amostras (raiva), 114 amostras (felicidade), 126 amostras (neutro) e 100 amostras (tristeza).

Utilizando $C = 20.0$ e os mesmos parâmetros de janela e passo dos testes iniciais, conseguimos obter uma taxa de reconhecimento de 93.6%. Isto, apenas duplicando as emoções com menos amostras. A matriz de confusão obtida é apresentada na tabela 6.

Tabela 6 – Matriz de confusão para SVM com quatro emoções após o balanceamento de amostras.

	Raiva	**Alegria**	**Neutro**	**Tristeza**
Raiva	19,30	3,62	0,01	0,00
Alegria	1,47	23,71	0,51	0,00
Neutro	0,00	0,09	27,80	0,55
Tristeza	0,00	0,00	0,17	22,77

O problema de discrepância entre quantidade de amostras claramente estava afetando a performance do classificador SVM, porém provavelmente esta precisão ainda pode ser melhorada. Então ajustamos os parâmetros de janela e passo para tentar aumentar a precisão obtida após o ajuste da quantidade de amostras de áudio.

5.4 Experimento Utilizando EMO-DB

Segundo Giannakopoulos (2015), os valores das janelas de curto prazo normalmente são de 20 a 100 milissegundos e sugerindo uma sobreposição de 50%. Já com as janelas de médio prazo ele relata que são utilizados os valores de 1 a 10 segundos e em alguns casos com sobreposição de até 90%.

Após a resolução do problema de desbalanceamento das amostras de áudio, prosseguimos com os experimentos utilizando SVM realizando um teste de força bruta que será descrito na próxima seção.

5.4.1 Ajuste de Parâmetros Utilizando o Método de Força Bruta

Para o ajuste dos parâmetros (janelas e passos), utilizamos uma estratégia de força bruta, onde foi realizado um treinamento com as janelas de médio prazo iniciando em 1000 ms, e aumentando 100 ms a cada iteração até atingir 3000 ms, utilizando 50% de sobreposição para cada janela. Em cada uma das iterações para a janela de médio prazo, foi ajustada a janela de curto prazo entre 20 e 100 ms, aumentando em 1 ms, e utilizando uma sobreposição de 33%.

Desta forma cobriu-se todas as combinações de parâmetros possíveis entre os valores:

- Janela de curto prazo: de 20, 21, 22, ..., 100 ms;
- Janela de médio prazo: de 1000, 1100, 1200, ..., 3000 ms.

A melhor acurácia obtida no teste acima descrito foi de 94,3%. Foi utilizado o banco de dados EMO-DB, utilizando-se 75% dos dados para treinamento e 25% para validação do modelo. Os valores ótimos encontrados foram de 36 ms para janela de curto prazo, e 1300 ms para a janela de médio prazo. Neste caso, o parâmetro de custo C que obteve o melhor desempenho foi 10,0. A matriz de confusão é apresentada na tabela 7.

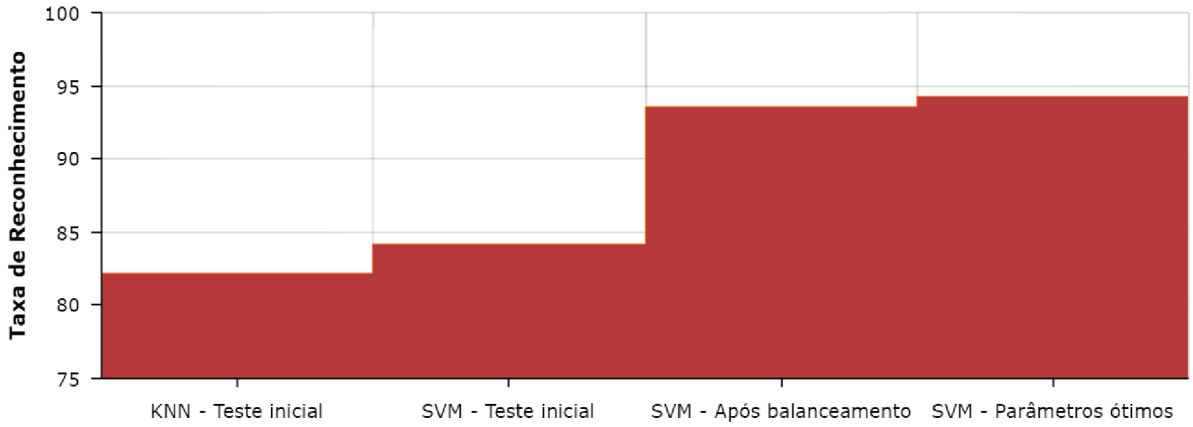
Tabela 7 – Matriz de confusão para SVM com quatro emoções após o ajuste de parâmetros.

	Raiva	**Alegria**	**Neutro**	**Tristeza**
Raiva	19,20	3,72	0,01	0,00
Alegria	1,08	24,29	0,31	0,00
Neutro	0,00	0,04	27,17	0,24
Tristeza	0,00	0,00	0,29	22,64

Podemos verificar que a taxa de reconhecimento é altíssima, a qual superou até mesmo o classificador de três estágios proposto por Iriya (2014), porém ele utilizou todas as emoções disponíveis no banco de dados, o que pode ter contribuído para a diminuição da performance do classificador proposto por ele.

A figura 9 exibe um gráfico *"step"* para fins de comparação dos resultados obtidos até o momento. É possível constatar que a grande melhora na taxa de reconhecimento ocorreu após o balanceamento das amostras entre as emoções.

Figura 9 Comparação entre os testes iniciais e os experimentos.



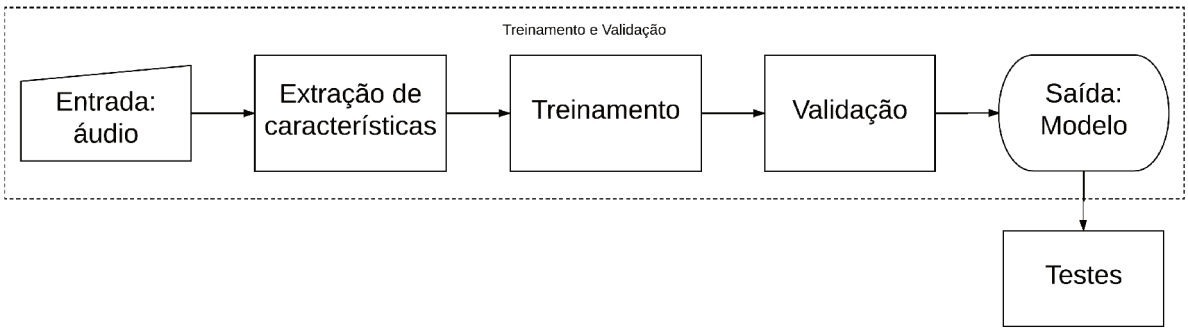
Fonte: Autoria própria.

5.4.2 Validação e Testes

A etapa de validação foi utilizada apenas para ajuste de parâmetros (*model tuning*), sendo que as amostras eram divididas pelo algoritmo, logo não seria possível separar os áudios que foram duplicados para esta etapa. Adicionalmente, a validação do modelo é implementada pela biblioteca criada por Giannakopoulos (2015) e, devido a falta de documentação de como a validação cruzada é realizada por esta biblioteca, decidimos adotar uma nova configuração do banco de dados, utilizando 60% das amostras para o treinamento, 20% para a validação, e 20% para uma nova etapa, a de teste. As amostras foram separadas aleatoriamente para cada uma das etapas.

Após esta definição, as etapas que são realizadas são ilustradas na figura 10, sendo que o conjunto de dados utilizado na etapa de testes não foi duplicado, portanto apresentará uma acurácia próxima à de uma situação real.

Figura 10 Etapas realizadas para validação e testes do modelo.



Fonte: Autoria própria.

Foi executado um novo teste, agora utilizando a acurácia da etapa de testes como resultado aceito, utilizando a etapa de validação somente para escolha do parâmetro de custo que corresponde ao melhor modelo.

Esta nova configuração foi testada utilizando os parâmetros ótimos obtidos no teste de força bruta:

- Janela de curto prazo: 36 ms com sobreposição de 33%;
- Janela de médio prazo: 1300 ms com sobreposição de 50%.

O parâmetro de custo (C) que obteve o melhor desempenho na etapa de validação (0.5) foi diferente do descrito anteriormente (10.0) devido à modificação na quantidade das amostras utilizadas para o treinamento.

Na etapa de testes, a acurácia diminuiu, como esperado, para 86,79% devido à utilização de amostras totalmente desconhecidas pelo modelo para se testar o mesmo. A matriz de confusão obtida é apresentada na tabela 8.

Tabela 8 – Matriz de confusão para SVM com quatro emoções na etapa de testes.

	Raiva	**Alegria**	**Neutro**	**Tristeza**
Raiva	80,0	20,0	0,00	0,00
Alegria	9,09	72,73	18,18	0,00
Neutro	0,00	0,00	100,0	0,00
Tristeza	0,00	0,00	0,00	100,0

É possível observar que as emoções neutro e tristeza obtiveram uma taxa de reconhecimento de 100%, o que é algo surpreendente para um classificador relativamente simples como o SVM.

5.5 Experimento Utilizando Banco de Dados em Português

Os testes iniciais utilizando o banco de dados em português foram realizados utilizando os mesmos parâmetros ótimos obtidos para o modelo com o EMO-DB. Este teste inicial foi realizado para comparação com os próximos testes, visando identificar o quanto uma linguagem influencia em um modelo de reconhecimento de emoções. Na etapa de validação, foi alcançado 49,2% de acurácia.

Pelo fato de a taxa de reconhecimento ter sido muito menor do que a esperada, realizamos um balanceamento da quantidade de amostras, o que resultou, na etapa de validação, em uma acurácia de 82,1%. A matriz de confusão obtida nesta etapa é exibida na tabela 9.

Tabela 9 – Matriz de confusão na etapa de validação utilizando BD em português.

	Raiva	**Alegria**	**Neutro**	**Tristeza**
Raiva	12, 71	4, 68	1, 35	1, 85
Alegria	2, 06	25, 76	1, 41	0, 18
Neutro	0, 94	1, 06	23, 65	0, 82
Tristeza	1, 47	0, 65	1, 41	20, 00

Seguindo-se o mesmo procedimento executado para o EMO-DB, realizamos o teste de força bruta como descrito na seção 5.4.1 visando encontrar parâmetros ótimos para o novo banco de dados testado, utilizando 60% dos dados para treinamento e 20% para validação. Os parâmetros ótimos encontrados neste teste foram de 100 ms para a janela de curto prazo e de 1200 ms para a janela de médio prazo. Foi encontrada uma acurácia de 84,4% na etapa de validação utilizando 20 como parâmetro de custo C.

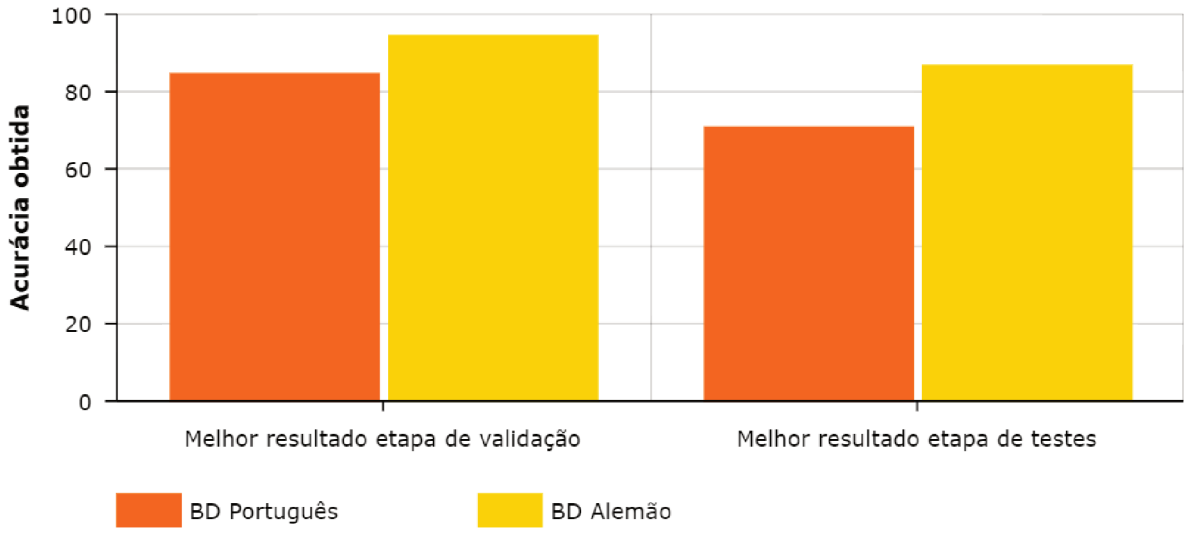
Tabela 10 – Matriz de confusão na etapa de testes utilizando BD em português.

	Raiva	**Alegria**	**Neutro**	**Tristeza**
Raiva	50, 0	37, 5	12, 5	0, 00
Alegria	0, 00	100, 0	0, 00	0, 00
Neutro	0, 00	20, 00	60, 0	20, 00
Tristeza	20, 00	0, 00	0, 00	80, 00

Os outros 20% das amostras foram utilizados para a etapa de testes, onde foi obtida uma acurácia de 70,83%. Observando a matriz de confusão da tabela 10, pode-se constatar que houve uma precisão de 100% para a emoção alegria e 80% para a emoção tristeza, o que é um bom resultado, considerando que os áudios não receberam um pré-processamento, o que poderia aumentar consideravelmente a precisão obtida.

Na figura 11, é apresentado graficamente uma comparação dos resultados entre os dois bancos de dados testados com a aplicação aqui desenvolvida.

Figura 11 Comparação entre os resultados obtidos com os dois bancos de dados.



Fonte: Autoria própria.

6 Conclusão

Com este trabalho foi possível alcançar o objetivo geral especificado, de reconhecer a emoção representada por um áudio entre um grupo de emoções pré-definidas as quais o algoritmo foi treinado, além de uma análise comparativa entre os classificadores SVM e KNN.

Antes de ser possível alcançar os resultados descritos, foi necessário adquirir bastante conhecimento teórico sobre o estado da arte em reconhecimento de emoções através da voz, onde chegamos à conclusão que não existe um consenso sobre qual método e técnicas são melhores para se obter um resultado e desempenho ótimo. Após isto, foi decidido utilizar os métodos mais simples, pois exigiam menor poder de processamento para o treinamento, pois era pretendido utilizar o método de força bruta para ajuste de parâmetros, o que foi realizado com êxito. Apesar de não ter sido mensurado o tempo exato de treinamento, o maior tempo não passou de poucos minutos, e o reconhecimento não passou de 1 segundos em nenhum dos testes realizados, resultado este que consideramos bastante aceitável para a maioria das aplicações.

Algumas limitações associadas ao desenvolvimento deste trabalho são: o tamanho do áudio a ser reconhecido, logo, audios longos podem não apresentar taxa condizente com os testes descritos; múltiplas emoções em um mesmo áudio; o regionalismo não foi considerado, podendo causar uma diminuição na taxa de reconhecimento.

Através de todos os testes realizados é possível observar a diferença da acurácia final obtida entre os dois banco de dados. A principal razão que encontramos se deve à origem das amostras de áudio. O banco de dados alemão EMO-DB têm uma qualidade muito superior, com mínimos ruídos, enquanto que os áudios em português tem uma qualidade muito inferior, com muitos ruídos e música ao fundo, por exemplo. Portanto, consideramos que o ponto fraco da abordagem utilizando os áudios em português, deve-se à qualidade dos mesmos.

A acurácia obtida de 86,79%, utilizando as amostras do EMO-DB, está dentro do esperado devido às características não terem sido selecionados adequadamente, onde foi utilizado somente as que eram extraídas pela biblioteca de treinamento. É possível

observar que as emoções neutro e tristeza obtiveram uma taxa de reconhecimento de 100%, o que é algo surpreendente para um classificador relativamente simples como o SVM.

Pode-se considerar que a precisão de 70,83% obtida foi bastante alta, salvo a baixa qualidade dos áudios em português. Com a realização de algum pré-processamento, e seleção das características mais adequadas a serem extraídas, pode-se esperar que a taxa de reconhecimento das emoções aumente consideravelmente.

Apesar de uma busca exaustiva por uma biblioteca de uso simples e fácil, a única que encontramos com suporte aos métodos que utilizam os algoritmos que se pretendeu realizar os testes, foi a *pyAudioAnalisis*. A principal limitação do sistema desenvolvido, está diretamente relacionada com esta biblioteca, que foi depender das características de áudios extraídas pela mesma. Talvez, se realizássemos testes com diferentes características, poderíamos selecionar as melhores e obter um resultado superior.

6.1 Trabalhos Futuros

Após a análise dos resultados encontrados para o reconhecimento com os áudios em português, pode-se chegar a conclusão de que seria necessário uma etapa de pré-processamento do áudio, antes da etapa de extração das características. Portanto, sugerimos o uma adição ao modelo desenvolvido neste trabalho, realizando-se um pré-processamento das amostras, o que certamente resultaria em uma melhora na taxa de reconhecimento. Uma aplicação para o pré-processamento do áudio, poderia ter aplicação, não somente para sistemas de reconhecimento de emoções, como para qualquer sistema que seja necessário uma boa qualidade de áudio.

Também seria interessante, realizar a mesma análise deste trabalho com todas as emoções disponíveis no banco de dados de Berlin, afim de comparar os resultados com alguns trabalhos do estado da arte que fizeram o reconhecimento utilizando todas estas emoções.

Tinha-se intuito inicial de utilizar o banco de dados do Centro de Atendimento a Emergências (190, 192, 193), porém não conseguimos a disponibilização dos dados, mas averiguamos que é possível realizar um projeto em parceria com o COPOM de Florianópolis. Um requisito adicional para o desenvolvimento desta aplicação para o COPOM é o reconhecimento em tempo real, pois ao receber uma ligação, o sistema deveria

informar ao atendente a emoção da pessoa com a qual está falando, sendo a identificação da emoção menos suscetível a erro se comparado ao julgamento humano.

Outra sugestão seria a captura de áudios de pessoas no seu cotidiano, que poderia ser utilizado para aproximar o reconhecimento de emoções à uma aplicação real, que pode ser a interação entre humano e computadores, por exemplo.

Uma forma de aumentar a taxa de reconhecimento para este sistema de reconhecimento de emoções, seria realizar uma seleção das características a serem extraídas, otimizando as informações obtidas através das mesmas, o que pode-se esperar que aumentaria razoavelmente a acurácia do sistema.

Referências

ALVA, M. Y.; NACHAMAI; PAULOSE, J. A comprehensive survey on features and methods for speech emotion detection. In: *2015 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT)*. [S.l.: s.n.], 2015. p. 1–6. Citado 4 vezes nas páginas 22, 37, 38 e 44.

BOERSMA, P. Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. 1993. Citado na página 39.

CHANDRASEKAR, P.; CHAPANERI, S.; JAYASWAL, D. Automatic speech emotion recognition: A survey. *IEEE International Conference On Circuits, Systems, Communication And Information Technology Applications (cscita)*, abr 2014. Disponível em: <http://dx.doi.org/10.1109/cscita.2014.6839284>. Citado 2 vezes nas páginas 30 e 32.

CHAPANERI, S.; JAYASWAL, D. Emotion recognition from speech using teager based dscc features. *International Journal Of Computer Applications*, out 2013. Citado na página 30.

CHAVHAN, Y.; DHORE, M. L.; YESAWARE, P. Speech emotion recognition using support vector machine. *International Journal Of Computer Applications*, v. 1, n. 20, p. 6–9, fev 2010. Citado na página 36.

CHEE, L. S. et al. Mfcc based recognition of repetitions and prolongations in stuttered speech using k-nn and lda. In: *2009 IEEE Student Conference on Research and Development (SCOReD)*. [S.l.: s.n.], 2009. p. 146–149. Citado na página 38.

GIANNAKOPOULOS. Pyaudioanalysis: An open-source python library for audio signal analysis. *Public Library of Science One*, v. 10, n. 12, dez 2015. Disponível em: <http://dx.doi.org/10.1371/journal.pone.0144610>. Citado 4 vezes nas páginas 46, 47, 56 e 58.

GUNES, H. et al. Emotion representation, analysis and synthesis in continuous space: A survey. In: *FG*. [S.l.: s.n.], 2011. Citado na página 26.

HARIMI, A. et al. Anger or joy? emotion recognition using nonlinear dynamics of speech. *Applied Artificial Intelligence*, Taylor & Francis, v. 29, n. 7, p. 675–696, 2015. Disponível em: <http://dx.doi.org/10.1080/08839514.2015.1051891>. Citado na página 27.

HOUWER, J. D.; HERMANS, D. *Cognition and Emotion Reviews of Current Research and Theories*. New York: Psychology Press, 2010. Citado na página 25.

IRIYA, R. *Análise de sinais de voz para reconhecimento de emoções*. Dissertação (Mestrado) — Curso de Engenharia e Sistemas Eletrônicos, Universidade de São Paulo, 2014. Citado 15 vezes nas páginas 11, 25, 26, 27, 28, 29, 32, 33, 34, 39, 40, 44, 46, 47 e 57.

JARANDE, S. S.; WAGHMARE, S. A survey on different classifier in speech recognition techniques. In: . [S.l.: s.n.], 2015. v. 5, n. 3. ISSN 2250-2459. Citado na página 36.

KISHORE, K. V. K.; SATISH, P. K. Emotion recognition in speech using mfcc and wavelet features. In: *2013 3rd IEEE International Advance Computing Conference (IACC)*. [S.l.: s.n.], 2013. p. 842–847. Citado na página 38.

LE, D.; PROVOST, E. M. Emotion recognition from spontaneous speech using hidden markov models with deep belief networks. In: *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*. [S.l.: s.n.], 2013. p. 216–221. Citado 4 vezes nas páginas 25, 34, 38 e 39.

LIN, Y.-L.; WEI, G. Speech emotion recognition based on hmm and svm. In: *2005 International Conference on Machine Learning and Cybernetics*. [S.l.: s.n.], 2005. v. 8, p. 4898–4901 Vol. 8. ISSN 2160-133X. Citado na página 33.

LIU, Y.; ZHENG, Y. One-against-all multi-class svm classification using reliability measures. *Proceedings. Ieee International Joint Conference On Neural Networks*, dez 2005. Disponível em: <http://dx.doi.org/10.1109/ijcnn.2005.1555963>. Citado na página 48.

LUGER, G. F. I. A. *6. ed.* São Paulo: Pearson Educação do Brasil, 2013. Citado na página 21.

MIERSWA, I.; MORIK, K. Automatic feature extraction for classifying audio data. *Machine Learning*, v. 58, n. 2-3, p. 127–149, fev 2005. Disponível em: <http://dx.doi.org/10.1007/s10994-005-5824-7>. Citado na página 32.

MILGRAM, J.; CHERIET, M.; SABOURIN, R. "one against one" or "one against all": Which one is better for handwriting recognition with svms? *Tenth International Workshop On Frontiers In Handwriting Recognition*, La Baule (France), out 2006. Citado na página 48.

NILOFER et al. Automatic emotion recognition from speech signals: A review. *International Journal Of Scientific & Engineering Research*, v. 6, n. 4, abr 2015. Citado 3 vezes nas páginas 28, 44 e 45.

NWE, T. L.; FOO, S. W.; SILVA, L. C. D. Speech emotion recognition using hidden markov models. *Speech Communication*, v. 41, n. 4, p. 603 – 623, 2003. ISSN 0167-6393. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0167639303000992>. Citado na página 33.

PAO, T. long; CHEN, Y. te; YEH, J. heng. Mandarin emotional speech recognition based on svm and nn. *18th International Conference On Pattern Recognition (icpr'06)*, 2006. Disponível em: <http://dx.doi.org/10.1109/icpr.2006.780>. Citado na página 34.

PEDREGOSA, F. et al. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, v. 12, n. Oct, p. 2825–2830, 2011. Citado na página 47.

SCHULLER, B.; RIGOLL, G.; LANG, M. Hidden markov model-based speech emotion recognition. In: IEEE. *Multimedia and Expo, 2003. ICME'03. Proceedings. 2003 International Conference on*. 2003. v. 1, p. I–401. Disponível em: <http://dx.doi.org/10-.1109/icme.2003.1220939>. Citado na página 46.

SUJATHA, B.; AMEENA, O. Speech emotion recognition using hmm and gmm and svm models. *International Journal Of Professional Engineering Studies (ijpes)*, p. 311–318, jul 2016. Citado na página 48.

VERVERIDIS, D.; KOTROPOULOS, C. Emotional speech recognition: Resources, features, and methods. *Speech communication*, Elsevier, v. 48, n. 9, p. 1162–1181, set 2006. Disponível em: <http://dx.doi.org/10.1016/j.specom.2006.04.003>. Citado 2 vezes nas páginas 43 e 44.

WHISSELL et al. A dictionary of affect in language: Iv reliability, validity, and applications. *Perceptual & Motor Skills*, v. 62, n. 3, p. 875–888, jun 1986. Disponível em: <http://dx.doi.org/10.2466/pms.1986.62.3.875>. Citado na página 26.

Apêndices

APÊNDICE A – Código fonte desenvolvido

```
#!/usr/bin/env python

import os
from collections import OrderedDict
from pyAudioAnalysis import audioTrainTest as aT
import argparse
import time
import sys
sys.path.append('C:\dev\emotion_speech_recognition\src\
    dependencies\pyAudioAnalysis')

def print_parameters(st_w, st_s, mt_w, mt_s, perc_train, use_svm,
    model_name):
    print "\n================ {} ================".format( "SVM"
        if use_svm else "KNN")
    print "Parameters for model '{}'".format(model_name)
    print
    print "Short term window: {}".format(st_w)
    print "Short term step: {}".format(st_s)
    print "Mid term window: {}".format(mt_w)
    print "Mid term step: {}".format(mt_s)
    print "Utilizing {:.0f}% of samples for training".format(
        perc_train * 100 )
    print "===================================="

def feature_and_train(samples_prefix, st_w, st_s, mt_w, mt_s,
    perc_train, confusion_matrix_perc, use_svm, verbosity,
    model_name):
     start = time.time()
```

```
24      list_of_dirs_or_classes = []
25      dirs = [d for d in os.listdir(samples_prefix) if os.path.
            isdir(os.path.join(samples_prefix, d))]
26      for dire in dirs:
27          list_of_dirs_or_classes.append(samples_prefix + str(dire)
                )
28
29      print "Starting training..."
30      if use_svm:
31          model_name = "models/SVM_"+model_name
32          print_parameters(st_w, st_s, mt_w, mt_s, perc_train,
                use_svm, model_name)
33          bestParam, bestAcc = aT.featureAndTrain(
                list_of_dirs_or_classes, mt_w, mt_s, st_w, st_s, "svm",
                 model_name, False, perc_train, confusion_matrix_perc,
                verbosity=verbosity)
34      else:
35          model_name = "models/KNN_"+model_name
36          print_parameters(st_w, st_s, mt_w, mt_s, perc_train,
                use_svm, model_name)
37          bestParam, bestAcc = aT.featureAndTrain(
                list_of_dirs_or_classes, mt_w, mt_s, st_w, st_s, "knn",
                 model_name, False, perc_train, confusion_matrix_perc,
                verbosity=verbosity)
38
39      end = time.time()
40      print "Finished in {:10.2f} seconds".format(end-start)
41      return bestAcc
42
43  def test_model(prefix, model):
44      dirs = os.listdir(prefix)
45      use_svm = True
46      model = ("models/SVM_" + model) if use_svm else ("models/KNN_
            " + model)
```

```
47
48      print "\nTesting model: {}".format(model)
49      print "Confusion Matrix:"
50      for d in dirs:
51          print "\t{}".format(d[:4]),
52      print
53
54      total_correct = 0
55      total_files = 0
56      for classs in dirs:
57          files = os.listdir(prefix+classs)
58          class_files_num = len(files)
59          total_files = total_files+class_files_num
60          classified = OrderedDict()
61          for dir in dirs:
62              classified[dir] = 0
63
64          print "{}\t".format(classs[:4]),
65          for file in files:
66              file_to_test = prefix+classs+"/"+file
67              class_classified = test_file(file_to_test, model,
                    use_svm)
68              classified[class_classified] = classified[
                    class_classified]+1
69              if class_classified == classs:
70                  total_correct = total_correct+1
71
72          for key in classified.keys():
73              print "{}\t".format( round((classified[key]/float(
                    class_files_num))*100,2) ),
74          print
75      print
76
77      print "General accuracy is {}%".format( round(total_correct/
```

```
        float(total_files)*100,2) )
78
79  def test_file(filename_to_test, model_name, use_svm=True, verbose
        =False):
80      if os.path.isfile(filename_to_test):
81          start = time.time()
82          if use_svm:
83              r, P, classNames = aT.fileClassification(
                    filename_to_test, model_name, "svm")
84          else:
85              r, P, classNames = aT.fileClassification(
                    filename_to_test, model_name, "svm")
86
87          chosen = 0.0
88          chosenClass = ""
89          if len(P) == len(classNames):
90              for i in range(0, len(P), 1):
91                  if P[i] > chosen:
92                      chosen = P[i]
93                      chosenClass = classNames[i]
94
95          end = time.time()
96          if verbose:
97              print "\n\nThe audio file was classified as {} with
                    prob {}% in {:.2f} seconds\n\n".format(chosenClass.
                    upper(), round(chosen*100, 2), end - start )
98          return chosenClass
99      else:
100         print "File doesnt exists: {}".format(filename_to_test)
101         return None
102
103 def train_until_get_better_acc(samples_prefix, model_name,
        train_until):
104     best_acc = 0
```

```
105     for i in range(0,1000):
106         accuracy = train_SVM(samples_prefix, model_name)
107         print "\nCurrent accuracy: {}".format(accuracy)
108         print "Best accuracy: {}\n\n".format(best_acc)
109 
110         if accuracy > best_acc:
111             prefix = samples_prefix+"../../src/models/"
112             os.rename(prefix+"SVM_port_single", prefix+"
                    best_SVM_port_single")
113             os.rename(prefix+"SVM_port_single.arff", prefix+"
                    best_SVM_port_single.arff")
114             os.rename(prefix+"SVM_port_singleMEANS", prefix+"
                    best_SVM_port_singleMEANS")
115             best_acc = accuracy
116 
117         if accuracy >= train_until:
118             break
119 
120 def train_SVM(samples_prefix, model_name):
121 
122     # ======= PORTUGUESE BEST CONFIGURATION ========
123     SHORT_TERM_WINDOW = 0.1
124     SHORT_TERM_STEP = 0.033
125     MID_TERM_WINDOW = 1.2
126     MID_TERM_STEP = 0.6
127     # ======= PORTUGUESE BEST CONFIGURATION ========
128 
129     # ======= GERMAN BEST CONFIGURATION ========
130     SHORT_TERM_WINDOW = 0.036
131     SHORT_TERM_STEP = 0.012
132     MID_TERM_WINDOW = 1.3
133     MID_TERM_STEP = 0.65
134     # ======= GERMAN BEST CONFIGURATION ========
135 
```

```
136     confusion_matrix_perc = True
137     use_svm = True
138     perc_train = 0.75
139     VERBOSITY = False
140 
141     return feature_and_train(samples_prefix, SHORT_TERM_WINDOW,
            SHORT_TERM_STEP, MID_TERM_WINDOW, MID_TERM_STEP,
142                                  perc_train,
                                        confusion_matrix_perc, use_svm
                                        , VERBOSITY, model_name)
143 
144 def train_KNN(samples_prefix, model_name):
145     SHORT_TERM_WINDOW = 0.036
146     SHORT_TERM_STEP = 0.012
147     MID_TERM_WINDOW = 1.3
148     MID_TERM_STEP = 0.65
149 
150     confusion_matrix_perc = True
151     use_svm = False
152     perc_train = 0.75
153     VERBOSITY = False
154 
155     feature_and_train(samples_prefix, SHORT_TERM_WINDOW,
            SHORT_TERM_STEP, MID_TERM_WINDOW, MID_TERM_STEP,
156                                  perc_train,
                                        confusion_matrix_perc, use_svm
                                        , VERBOSITY, model_name)
157 
158 def brute_force_training(samples_prefix):
159     min_st    = 0.020
160     max_st    = 0.100
161     step_st   = 0.001
162     st_overl = 0.33
163 
```

```
164     min_mt   = 1.000
165     max_mt   = 3.000
166     step_mt  = 0.100
167     mt_overl = 0.5
168
169     MID_TERM_WINDOW = min_mt
170     MID_TERM_STEP = round(MID_TERM_WINDOW*mt_overl, 3)
171     SHORT_TERM_WINDOW = min_st
172     SHORT_TERM_STEP = round(SHORT_TERM_WINDOW*st_overl, 3)
173
174     confusion_matrix_perc = True
175     use_svm = True
176     perc_train = 0.75
177     VERBOSITY = False
178
179     bestAcc = 0.0
180     bestAccParams = {
181         "st_w": SHORT_TERM_WINDOW,
182         "st_s": SHORT_TERM_STEP,
183         "mt_w": MID_TERM_WINDOW,
184         "mt_s": MID_TERM_STEP
185     }
186
187     range_mt_max = int( round(max_mt - min_mt, 3) / step_mt)+1
188     range_st_max = int( round(max_st - min_st, 3) / step_st)+1
189
190     for mt in range(0,range_mt_max):
191         SHORT_TERM_WINDOW = min_st
192         SHORT_TERM_STEP = round(SHORT_TERM_WINDOW*st_overl, 3)
193         for st in range(0,range_st_max):
194             accuracy = feature_and_train(samples_prefix,
                    SHORT_TERM_WINDOW, SHORT_TERM_STEP, MID_TERM_WINDOW
                    ,
195                      MID_TERM_STEP, perc_train,
```

```
                        confusion_matrix_perc, use_svm, VERBOSITY,
                        "brute_force_test_model")
196
197             SHORT_TERM_WINDOW = SHORT_TERM_WINDOW +step_st
198             SHORT_TERM_STEP = round(SHORT_TERM_WINDOW*st_overl,
                    3)
199
200             if accuracy > bestAcc:
201                 bestAcc = accuracy
202                 bestAccParams["st_w"] = SHORT_TERM_WINDOW
203                 bestAccParams["st_s"] = SHORT_TERM_STEP
204                 bestAccParams["mt_w"] = MID_TERM_WINDOW
205                 bestAccParams["mt_s"] = MID_TERM_STEP
206
207         MID_TERM_WINDOW = MID_TERM_WINDOW+step_mt
208         MID_TERM_STEP = round(MID_TERM_WINDOW*mt_overl, 3)
209
210         print "\n\n\nMelhor precisao: {}".format(bestAcc)
211         print "SHORT_TERM_WINDOW = {} \nSHORT_TERM_STEP = {} \
                nMID_TERM_WINDOW = {} \nMID_TERM_STEP = {}\n\n".format(
212             bestAccParams["st_w"], bestAccParams["st_s"],
                    bestAccParams["mt_w"], bestAccParams["mt_s"])
213
214
215
216 if __name__ == '__main__':
217     PROJECT_PATH = os.path.dirname(os.path.abspath(__file__))+"
            /../"
218
219     parser = argparse.ArgumentParser()
220     parser.add_argument("--db", choices=['portuguese', 'german'],
            help="Which audio DB to use.")
221     parser.add_argument("--train", help="Train and validate the
            model.", action="store_true")
```

```
222     parser.add_argument("--test", help="Test the model with pre
            defined testset.", action="store_true")
223     parser.add_argument("--stest", help="Test model with a single
            file.", action="store_true")
224     parser.add_argument("--filename", help="The filename to test.
            ")
225     args = parser.parse_args()
226
227     MODEL_NAME = ""
228     if args.db == "german":
229         SAMPLES_PREFIX = PROJECT_PATH + 'audio_samples/german_emo
                /'
230         TEST_PREFIX = PROJECT_PATH + 'audio_samples/
                german_emo_test/'
231         MODEL_NAME = "german_single"
232     else:
233         SAMPLES_PREFIX = PROJECT_PATH + 'audio_samples/portuguese
                /'
234         TEST_PREFIX = PROJECT_PATH + 'audio_samples/
                portuguese_test/'
235         MODEL_NAME = "port_single"
236
237     if args.stest:
238         if not args.filename:
239             print("You must to specify the full path to the audio
                    file.")
240         else:
241             test_file(args.filename, "models/SVM_"+MODEL_NAME,
                    verbose=True)
242     elif args.db:
243         if args.train:
244             if args.test:
245                 parser.error("You can't use test and train flag
                        at the same time!")
```

```
246             else:
247                 brute_force_training(SAMPLES_PREFIX)
248                 train_until_get_better_acc(SAMPLES_PREFIX,
                        MODEL_NAME, 85)
249                 train_SVM(SAMPLES_PREFIX, MODEL_NAME)
250                 train_KNN(SAMPLES_PREFIX, MODEL_NAME)
251         elif args.test:
252             test_model(TEST_PREFIX, MODEL_NAME)
253 
254         else:
255             parser.error("You should specify to train or test
                with flags --train, --test or --stest")
256     else:
257         parser.error("You should specify which audio db to use
            with flag --db portuguese or --db german")
```

APÊNDICE B – Artigo SBC sobre o TCC

Reconhecimento Automático de Emoções Através da Voz

Jair da Rosa Júnior[1]

[1]Sistemas de Informação – Departamento de Informática e Estatística
Universidade Federal de Santa Catarina (UFSC) – Florianópolis, SC – Brazil

jdarosaj@gmail.com.br

Abstract. *After the emergence of the phones and more recently the computers, it become possible storing audios in digital format. Modern cell phones along with the internet, have made recording and transmitting these audios on a large scale viable. Then, a new demand of processing and extracting information arises. The speech emotion recognition is a recent demand, which only appeared because of the popularization of machine learning algorithms, where stands out KNN, SVM, GMM and HMM. In this work, we propose a SVM-based system, where voice features are extracted (like energy and pitch) and a supervised model is trained, utilizing each emotion to be recognized as a class. The recognition is given by the class with the highest likelihood. Using the Berlin Database of Emotional Speech (Emo-DB), we achieve a recognizing rate of 86,79% and using a Portuguese database, we've reached a rate of 70,83%. The obtained results were very reasonable, since some authors of the state-of-the-art got worse results.*

Resumo. *Após o surgimento dos telefones, e mais recentemente dos computadores, se tornou possível o armazenamento de áudios no formato digital. Os celulares modernos juntamente com a internet tornaram viável a gravação e transmissão destes áudios em larga escala. Surge então uma nova demanda de processamento e extração de informação dos mesmos. O reconhecimento de emoções através da voz é uma demanda recente, que só apareceu com a popularização de algoritmos de aprendizado de máquina, onde se destacam KNN, SVM, GMM e HMM. Neste trabalho foi proposto um sistema baseado em SVM, onde são extraídas características da voz (tais como pitch e energia) e um modelo é treinado de forma supervisionada, utilizando cada emoção a ser reconhecida como uma classe. O reconhecimento se dá, pela classe com maior verossimilhança obtida.*

Utilizando o banco de dados emocional de Berlin (em alemão) conseguimos obter uma taxa de reconhecimento de 86,79% e com o banco de dados criado em português, extraindo-se trechos de filmes e vídeos, foi obtida uma taxa de 70,83%. Os resultados obtidos foram bastante razoáveis, visto que alguns autores do estado da arte obtiveram resultados piores.

1. Introdução

A transmissão de voz por sistemas eletrônicos foi uma revolução inimaginável na forma como nos comunicamos e já existe há mais de um século. Com a popularização dos telefones e mais recentemente com a dos computadores, a integração entre telefonia e computação se tornou algo que certamente iria acontecer.

Essencialmente, a fala serve para transmitir uma mensagem através de palavras. Contudo, ela pode transmitir muito mais do que apenas palavras, pois possui características intrínsecas à ela – sonoridade, passo, entonação, nitidez, articulação, irregularidade, instabilidade e velocidade de fala são algumas delas. Com a análise destas características por algoritmos de Inteligência Artificial (IA) torna-se possível o reconhecimento automático da emoção do interlocutor.

IA é um conceito que não tem uma definição exata. Pelo fato da inteligência em si não ser um conceito bem definido, vários autores conceituam o termo de diferentes formas. Segundo Luger (2013), a IA pode ser definida como o ramo da ciência da computação que se ocupa da automação do comportamento humano.

Alva, Nachamai e Paulose (2015) descrevem que computação emocional é uma área da inteligência artificial que busca preencher o gap entre emoções humanas e tecnologia da computação. Por estudar emoções, logicamente, este campo de pesquisa é multidisciplinar, envolvendo estudos das áreas de ciência da computação, ciência cognitiva e principalmente da psicologia.

O principal motivo para o desenvolvimento deste projeto é o fato de ainda não existir um sistema de reconhecimento de emoções que funcione com uma precisão aceitável e de forma genérica. Um sistema com tal precisão, pode ser bastante útil para reconhecer emoções em ligações telefônicas, em um call center por exemplo, visando identificar se um cliente está estressado ou nervoso com quem está lhe atendendo, podendo indicar a qualidade do atendimento. Ou ainda, o reconhecimento de emoções durante um atendimento da central de emergência da polícia, onde a emoção da pessoa atendida poderia auxiliar em uma tomada de decisão.

2. Emoções

Emoções fazem parte do cotidiano, porém poucas pessoas sabem a fundo o que significa a palavra emoção. A fim de compreender o que é uma emoção, é necessário o estudo da teoria das emoções.

É necessário entender que o significado da palavra emoção é bastante amplo e, muitas vezes, é confundido com as reações do corpo ao vivenciar um estado emotivo. Estas reações são de suma importância para o reconhecimento automático de emoções, visto que esta tarefa consiste em avaliar as alterações na voz para concluir qual o estado emocional de um indivíduo.

Segundo Le e Provost (2013), a expressão da emoção é um processo dinâmico e complexo, portanto, o estudo das emoções envolve várias áreas, mas principalmente a psicologia, onde existem várias teorias que explicam a manifestação da emoção (HOUWER; HERMANS, 2010 apud IRIYA, 2014). As teorias de emoções mais conhecidas são a de James-Lang e de Cannon-Bard.

James-Lang propôs que um indivíduo, após receber um estímulo exterior, sofre alterações fisiológicas perturbadoras, sendo o reconhecimento desses sintomas pelo cérebro o que gera a emoção. Já a Teoria de Cannon-Bard sugere que as reações fisiológicas e a emoção ocorrem simultaneamente, uma vez que a interpretação do estímulo exterior ocorre em duas partes diferentes do cérebro. Há ainda algumas outras teorias, como as cognitivistas, que afirmam que os processos cognitivos, como percepções e recordações, são fundamentais para se perceberem as emoções. Tendo como exemplo a teoria de SchachterSinger, que presume que a experiência da emoção cresce a partir da consciência de excitação fisiológica (IRIYA, 2014).

Existem também algumas outras teorias de emoções citadas no survey realizado por Gunes et al. (2011). Estas teorias são distintas, porém todas afirmam que emoções possuem uma causa externa ao corpo humano, que geram alterações fisiológicas e que podem ser diretamente observadas. Estas afirmações são de vital importância para o desenvolvimento de um sistema de reconhecimento automático de emoções.

3. Características de Voz

As emoções devem ser modeladas e reconhecidas pelas alterações fisiológicas da voz humana, quando um indivíduo vivencia um determinado estado emocional. Estas alterações podem envolver tanto a prosódia - o estudo do ritmo, entonação e demais atributos correlatos na fala - quanto a qualidade da voz, relacionada à inteligibilidade e à naturalidade da fala.

Apesar de não haver uma regra de quais características utilizar para o reconhecimento, vários autores utilizam algumas em comum. As principais características prosódicas utilizadas são o pitch, que pode ser entendido como a altura (frequência) percebida num sinal de audio; a energia (intensidade sonora percebida pelo ouvido humano) e propriedades relacionadas à duração da voz e pausas, enquanto que entre as características que descrevem a qualidade da voz encontram-se as frequências formantes, que representam picos na resposta em frequência do aparelho fonador humano; a distribuição espectral, representada por Mel Frequency Cepstral Coefficients, que são uma representação paramétrica do espectro de frequências do sinal de voz; ou Log Frequency Power Coefficients, que são coeficientes que refletem a distribuição espectral de energia do sinal; a relação harmônicos/ruído (Harmonic to Noise Ratio ou HNR) e o fluxo glotal.

Algumas características podem ter maior importância que outras, ou ainda podem ser consideradas redundantes. Considerando isto, a seleção de características mais importantes diminui a complexidade computacional. Na prática, o excesso delas também pode ser prejudicial no desempenho do sistema pois o conjunto de dados de entrada é finito e pode apresentar tendência.

4. Métodos de Classificação

O reconhecimento de emoções através da voz tem sido resolvido como um problema estatístico ou de reconhecimento de padrões, onde um modelo é gerado extraindo-se características de voz dos áudios, cujas emoções são conhecidas e posteriormente uma instância cuja emoção é desconhecida é confrontada com os modelos existentes e o modelo mais adequado é escolhido. Os parâmetros de voz podem ser globais ou dinâmicos. Estes dois tipos deram origem à dois tipos de classificação: estática, que utiliza parâmetros de médio prazo e dinâmica que utiliza o conjunto dos parâmetros de curto prazo.

4.1. Algoritmos de Classificação

A seguir serão apresentados os métodos de classificação mais populares e que alguns autores já confirmaram que apresentam bons resultados para o tipo de classificação que pretendemos implementar neste trabalho, como o trabalho realizado por Le e Provost (2013) utilizando HMM e o sistema para reconhecimento de mandarim realizado por Pao, Chen e Yeh (2006) que alcançou incríveis 84.2% de precisão.

4.1.1. Modelos de Misturas de Gaussianas

Modelos de Misturas de Gaussianas (GMM) são um tipo particular de modelos de misturas, cuja importância é indiscutível para a área de processamento de voz e principalmente para os temas de reconhecimento de voz e de locutor.

GMM’s são largamente utilizados para estimar funções densidade de probabilidade desconhecidas, onde a tarefa de estimação torna-se a de estimar os pesos de cada gaussiana e suas respectivas médias e matrizes de covariância a partir dos dados observados.

4.1.2. Modelos Ocultos de Markov

Modelos Ocultos de Markov (HMM) apresentam um grande potencial para modelar as emoções humanas, uma vez que possuem o poder de manter informações temporais da sequência de observações através do uso de estados.

HMM’s são modelos estatísticos, associados ao Processo de Markov e às Cadeias de Markov. Neles, a ocorrência de uma sequência de eventos observáveis é decorrente de os eventos terem percorrido uma sequência de estados, sendo que cada estado emite um símbolo ou conjunto de símbolos observáveis, segundo alguma distribuição probabilística.

O Processo de Markov é um processo estocástico controlado por uma variável aleatória que representa o estado do sistema, cujo valor futuro depende dos estados passados. No processo de Markov de primeira ordem, o valor futuro depende apenas do estado atual, não da sequência de estados que o procederam, propriedade conhecida como Propriedade de Markov.

4.1.3. K-Vizinhos Mais Próximos

O K-Vizinhos Mais Próximos (KNN) é um algoritmo não-paramétrico utilizado em reconhecimento de padrões com parâmetros globais. É um dos algoritmos mais simples utilizados em aprendizagem de máquina, no qual não são gerados modelos propriamente ditos.

Em vez disso, a fase de treinamento consiste simplesmente em armazenar o vetor de parâmetros das amostras de treinamento e suas conhecidas classes. Uma amostra de teste cuja classe é desconhecida, é então classificada de acordo com sua proximidade às amostras de treinamento, que são chamadas de vizinhos. A classe em que a amostra será classificada é aquela que se repete mais vezes dentre os vizinhos mais próximos.

4.1.4. Máquinas de Vetores de Suporte

Máquinas de Vetores de Suporte (SVM) é uma técnica utilizada principalmente para reconhecimento de padrões, onde segundo Chavhan, Dhore e Yesaware (2010), normalmente é utilizado como classificador binário, porém também pode ser utilizado como um classificador multiclasses.

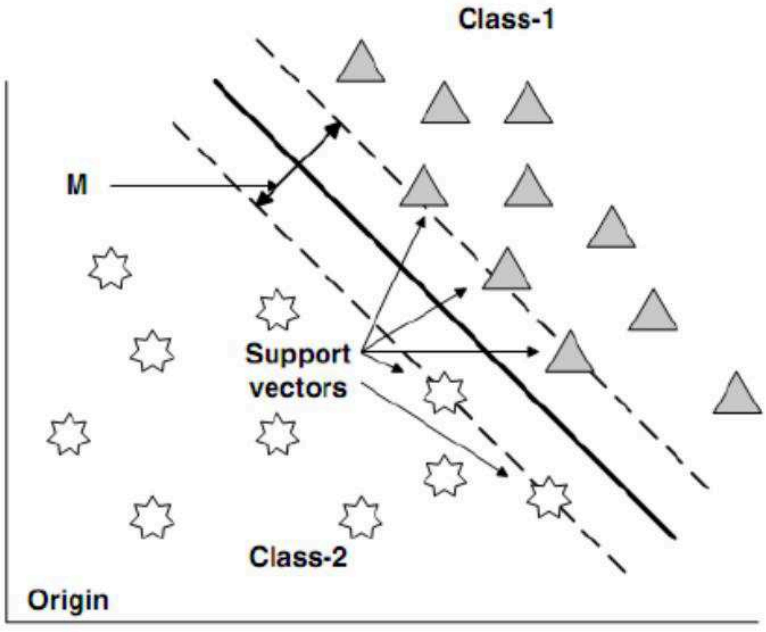


Figura 1. – Separação de duas classes com SVM

Segundo Jarande e Waghmare (2015), o intuito básico de SVM é criar um hiperplano, onde seu objetivo é separar as entradas em duas classes. A margem M é a distância entre os dois pontos mais próximos das duas diferentes classes. O classificador SVM posiciona a borda de decisão utilizando a margem máxima entre todos os possíveis hiperplanos.

5. Proposta

O sistema proposto contempla 2 módulos: módulo de entrada (extração de características e geração e validação do modelo) e um módulo de classificação (identificação da emoção).

O sistema de reconhecimento automático de emoções proposto pode ser construído de diversas formas, porém a intenção é desenvolvê-lo de forma que obtenha a maior precisão possível. Para isto se tornar possível é necessário entender como funciona cada um dos métodos de classificação, quais as técnicas utilizadas, e quanto o conjunto de emoções influencia na taxa de reconhecimento, bem como quais os métodos de classificação já existem implementados, pois não é escopo deste trabalho implementar algoritmos de classificação ou extração de características. Um esquema de como se pretende desenvolver este sistema para reconhecimento automático de emoções é apresentado na figura 2.

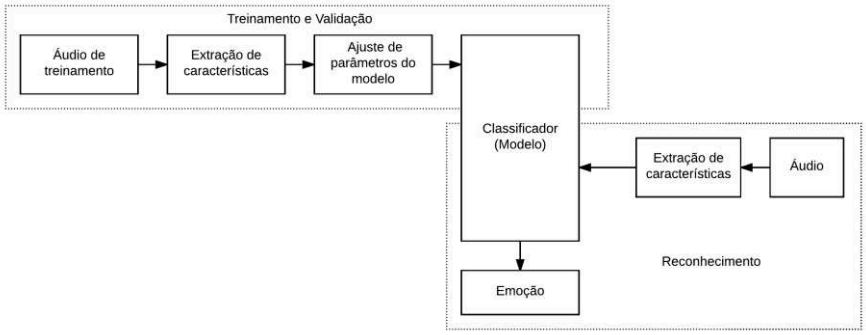


Figura 2. – Principais etapas do sistema proposto.

5.1. Emoções

Vários autores utilizam conjuntos de emoções ligeiramente diferentes. No estudo realizado por Ververidis e Kotropoulos (2006) as emoções mais comuns que aparecem nos banco de dados pesquisados são: Raiva (Anger), Alegria (Hapiness, Joy), tristeza (Sadness), Medo (Fear), Surpresa (Surprise) e estresse (Stress). Porém, alguns destes bancos de dados nem mesmo foram criados com o propósito de reconhecimento de emoções, e podem não ser adequados para este uso.

Neste artigo, foi desenvolvido um sistema para reconhecer um conjunto limitado de emoções, visando aumentar a precisão geral do reconhecimento. Pois, segundo Iriya (2014), deve-se definir um conjunto mais limitado possível somente com as emoções mais básicas, que sejam de fácil identificação do ponto de vista humano, e principalmente àquelas cujo reconhecimento automático tenha aplicação prática. Portanto, serão utilizadas apenas 4 emoções: Neutro, Raiva, Felicidade e Tristeza.

5.2. Banco de Dados de Áudio

Tendo em vista que um dos objetivos específicos deste trabalho, descrito na seção 1.2.1, é desenvolver a aplicação para a língua portuguesa e não foi obtido êxito na

procura de bancos de dados adequados para este fim nesta linguagem, optou-se por extrair trechos de áudios de filmes e vídeos brasileiros manualmente, visando cumprir este objetivo. Mesmo com a base em Português, foi utilizado também o banco de dados de emoções de Berlin (EMO-DB) para realizar os testes iniciais, pois é um banco de dados bastante estável, utilizado em diversos estudos como o de Iriya (2014), e o de Alva, Nachamai e Paulose (2015).

5.2.1. Banco de Dados Áudios de Emoções de Berlin

O banco de dados EMO-DB é atuado, ou seja, não foi capturado em uma situação em que a emoção realmente aconteceu, mas foi gravado por atores em um estúdio. Ele tem aproximadamente 530 áudios, com 10 frases diferentes faladas em alemão, sendo que estas frases não possuem nenhum significado que influencie no reconhecimento da emoção. Estas frases foram faladas por 5 locutoras mulheres e 5 locutores homens.

No banco de dados são considerados o estado neutro (63 amostras) mais 6 emoções, sendo elas raiva (101 amostras), tédio (65 amostras), nojo (37 amostras), ansiedade/medo (55 amostras), felicidade (57 amostras) e tristeza (50 amostras). Para os nossos testes, foram utilizados apenas as emoções raiva, felicidade, tristeza e neutro deste banco de dados pois correspondem às emoções dos áudios obtidas em português. Também balanceamos a quantidade de amostras que existem para cada uma das emoções, deixando cada emoção com a mesma quantidade de amostras de áudio.

5.2.2. Banco de Dados em Português

A criação do banco de dados na língua portuguesa (disponível em https://github.com/jdarosaj/emotion_portuguese_database) foi realizada extraindo-se trechos de filmes ou vídeos do YouTube3 com apenas uma frase e poucos segundos de duração. Todos os áudios foram extraídos utilizando o software Audacity.

A classificação foi realizada de forma manual pelo autor deste trabalho, através da análise dos trechos de áudio extraídos e escolha da emoção que apresentava a maior similaridade com as emoções do escopo deste trabalho. Devido à separação em classes destes áudios ter sido feita de forma manual, poderá haver uma taxa de erro associado à esta classificação.

O banco de dados contém 37 áudios que representam a emoção raiva, 27 a emoção felicidade, 22 a emoção tristeza e 24 áudios que não expressam emoções, sendo considerados neutros, completando 110 amostras no total.

5.3. Extração de Características

O sistema desenvolvido neste trabalho, faz uso da biblioteca pyAudioAnalisys5 criada por Giannakopoulos (2015), que utilizou outra biblioteca desenvolvida por Pedregosa et al. (2011) chamada *scikit-learn*, para o treinamento de SVM. A pyAudioAnalisys realiza a extração das características do áudio apresentados na tabela 1.

Tabela 1. Características extraídas pela biblioteca *pyAudioAnalisys*.

NOME DA CARACTERÍSTICA	DESCRIÇÃO
Taxa *Zero Crossing*	A taxa de mudança do sinal durante a duração de um frame em específico.
Energia	A soma dos quadrados dos valores do sinal, normalizados pelo respectivo comprimento do *frame*.
Entropia da Energia	A entropia das energias normalizadas dos sub-frames. Pode ser interpretado como uma medida de mudanças rápidas.
Centróide Espectral	O centro de gravidade do espectro.
Espalhamento do Espectro	O segundo momento central do espectro.
Entropia Espectral	Entropia das energias espectrais normalizadas para um conjunto de *sub-frames*.
Fluxo Espectral	A diferença quadrática entre as magnitudes normalizadas do espectro de dois *frames* sucessivos.
Deslocamento espectral	A frequência abaixo da qual se concentra 90% da distribuição de magnitude do espectro.
MFCC	Os coeficientes cepstrais de frequência Mel formam uma representação cepstral onde as bandas de frequência não são lineares, mas distribuídas de acordo com a escala Mel.
Vetor de *Chroma*	Uma representação da energia espectral com 12 elementos em que os recipientes representam 12 classes de *pitch* com comportamento constante de músicas ocidentais (espaçamento de semitom).
Desvio de *Chroma*	O desvio padrão dos 12 coeficientes de *chroma*.

5.3. Classificadores

Optou-se pela utilização de KNN e SVM para os testes iniciais, por serem classificadores simples, amplamente implementados, fáceis de encontrar, e que não necessitam de muita capacidade de processamento, se comparados à HMM e GMM. Esta decisão também foi influenciada pela escolha de outros autores, como Iriya (2014), já validaram a eficácia de HMMs e GMMs, mas principalmente por influência do trabalho de Sujatha e Ameena (2016) onde é mostrado que, em reconhecimento de emoções dependente do texto, o SVM é 15% a 20% superior aos classificadores citados anteriormente, verificaremos se o reconhecimento independente de texto também apresenta um bom resultado.

O classificador SVM é, essencialmente, utilizado para classificação binária, mas existem implementações para classificação multiclasse. Na próxima seção serão sucintamente abordadas as duas principais técnicas utilizadas na literatura.

6. Experimentos

Para cada um dos modelos treinados utilizamos várias configurações de parâmetros, e a cada rodada de treinamento/testes é calculada uma matriz de confusão para os testes de validação. Esta matriz permite a visualização do desempenho do algoritmo, onde cada linha da matriz representa as instâncias de uma classe predita, enquanto que cada coluna representa as instâncias da classe real. Neste trabalho apresentaremos em porcentagem ao invés de valores absolutos, visando facilitar sua interpretação.

Para definirmos qual o parâmetro que resultou em um melhor treinamento, é utilizada a métrica de acurácia, sendo esta a proporção de predições corretas, que é obtida através do seguinte cálculo:

$$(\sum d / \sum e) * 100$$

Onde d é cada elemento da diagonal da matriz de confusão, que representam os acertos. Já o e representa cada um dos elementos da matriz de confusão representando todo o conjunto de testes. O resultado deste cálculo será um número de 0 a 100 representando a porcentagem de acurácia do modelo.

5.2.1. Testes Iniciais

Os primeiros testes foram realizados apenas com o Berlin Emo-DB visto que este banco de dados já foi extensivamente testado e utilizado por diversos autores, sendo este bastante estável. Foram realizados testes com diferentes classificadores a fim de encontrar qual obteve a melhor performance. Todos os testes iniciais foram realizados utilizando 75% dos áudios disponíveis (separados aleatoriamente) para treinamento dos modelos e os outros 25% para testar os modelos gerados com os diversos parâmetros.

O primeiro teste com KNN foi realizado utilizando todas as emoções disponíveis no banco de dados EMO-DB. O parâmetro que obteve a melhor acurácia foi $K = 3$, com 66.7% de acurácia, resultando na matriz de confusão apresentada na tabela 2.

Tabela 2. Matriz de confusão para KNN com todas as emoções.

	Raiva	Tédio	Nojo	Medo	Alegria	Neutro	Tristeza
Raiva	21,26	0,02	0,44	0,03	1,77	0,06	0,00
Tédio	0,08	10,42	0,25	0,23	0,01	3,26	0,84
Nojo	1,22	0,32	4,91	0,97	0,13	0,72	0,23
Medo	2,18	1,38	2,25	5,81	0,89	0,58	0,11
Alegria	4,47	0,73	0,79	0,86	5,92	0,43	0,00
Neutro	0,09	4,62	0,29	0,16	0,32	8,74	0,87
Tristeza	0,00	1,11	0,26	0,27	0,00	0,05	9,62

O segundo teste foi realizado utilizando somente as emoções que pretendemos utilizar nos áudios em português, que são alegria, tristeza, raiva e neutro. Os parâmetros utilizados foram os mesmos do teste anterior. O melhor parâmetro K obtido neste segundo teste também foi $K = 3$, onde foi obtido 82.2% de acurácia. A matriz de confusão que foi obtida é exibida na tabela 3.

Tabela 3. Matriz de confusão para KNN com quatro emoções.

	Raiva	Alegria	Neutro	Tristeza
Raiva	34,07	2,88	0,36	0,00
Alegria	7,84	11,48	1,58	0,00
Neutro	0,33	1,04	19,81	2,70
Tristeza	0,00	0,00	1,07	16,84

A partir desses testes realizados é possível validar o que foi estudado anteriormente. Quanto maior for o universo de classes a ser reconhecido, maior a complexidade do sistema e consequentemente, menor a acurácia. Com isso se confirma a hipótese que tínhamos, de utilizar um conjunto limitado de emoções para não haver muito impacto na acurácia do modelo.

O primeiro teste com SVM foi realizado utilizando os mesmos valores de janela e passo do teste para KNN, a fim de se fazer uma comparação entre os dois classificadores. A melhor acurácia obtida foi de 71.5%. A matriz de confusão para este teste é apresentada na tabela 4.

Tabela 4. Matriz de confusão para SVM com todas as emoções.

	Raiva	Alegria	Neutro	Tristeza
Raiva	34,07	2,88	0,36	0,00
Alegria	7,84	11,48	1,58	0,00
Neutro	0,33	1,04	19,81	2,70
Tristeza	0,00	0,00	1,07	16,84

Na figura 3 é apresentada graficamente uma comparação da taxa de acerto entre os dois classificadores testados.

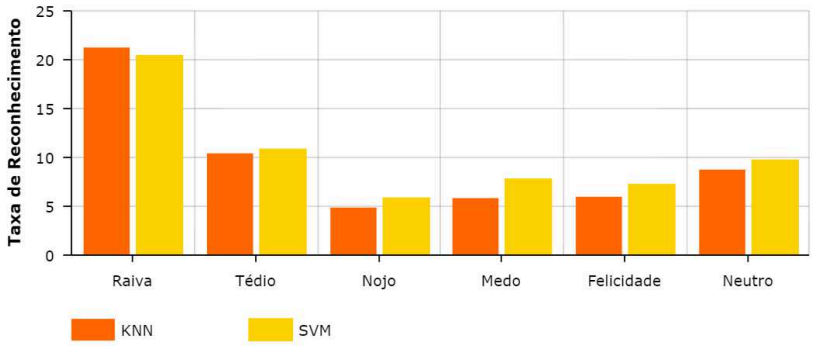


Figura 3. – Comparação da taxa de reconhecimento entre KNN e SVM.

No segundo teste realizado, utilizamos os mesmos parâmetros, porém utilizamos os áudios somente das 4 emoções que pretendemos reconhecer.

As emoções felicidade e neutro obtiveram melhor desempenho com o classificador SVM, enquanto que na emoção tristeza o desempenho foi similar, sendo o KNN ligeiramente superior apenas na emoção raiva. A taxa de reconhecimento deste modelo foi de 84.1%, contra 82.2% no método KNN. Podemos perceber que o gap entre

a precisão de um método e outro diminuiu bastante, apesar de SVM ainda ser mais preciso do que do método anterior.

Nos próximos testes utilizaremos apenas a classificação com SVM, por apresentar melhor precisão geral se comparado ao KNN.

5.2.2. Teste Final

Após alguns experimentos percebemos que a quantidade de amostras de cada uma das emoções eram muito discrepantes. Para resolver este problema tínhamos duas alternativas:

- Remover alguns áudios das emoções com mais amostras, com o mesmo objetivo anterior;
- Duplicar os áudios das emoções com menos amostras, visando equilibrar o número de áudios entre todas as emoções.

Optamos por duplicar os áudios das emoções com menos amostras, a fim de não perder nenhuma informação que possa ser útil ao modelo, pagando o preço de a quantidade de amostras continuar levemente desbalanceada.

Para o ajuste dos parâmetros (janelas e passos), utilizamos uma estratégia de força bruta, onde foi realizado um treinamento com as janelas de médio prazo iniciando em 1000 ms, e aumentando 100 ms a cada iteração até atingir 3000 ms, utilizando 50% de sobreposição para cada janela. Em cada uma das iterações para a janela de médio prazo, foi ajustada a janela de curto prazo entre 20 e 100 ms, aumentando em 1 ms, e utilizando uma sobreposição de 33%.

A figura 4 exibe um gráfico “step” para fins de comparação dos resultados obtidos até o momento. É possível constatar que a grande melhora na taxa de reconhecimento ocorreu após o balanceamento das amostras entre as emoções.

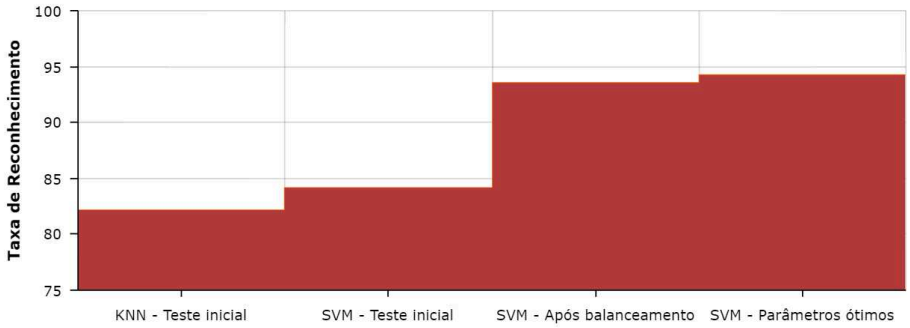


Figura 4. – Comparação entre os testes iniciais e os experimentos.

A etapa de validação foi utilizada apenas para ajuste de parâmetros (model tuning), sendo que as amostras eram divididas pelo algoritmo, logo não seria possível separar os áudios que foram duplicados para esta etapa. Adicionalmente, a validação do modelo é implementada pela biblioteca criada por Giannakopoulos (2015) e, devido a falta de documentação de como a validação cruzada é realizada por esta biblioteca, decidimos adotar uma nova configuração do banco de dados, utilizando 60% das

amostras para o treinamento, 20% para a validação, e 20% para uma nova etapa, a de teste. As amostras foram separadas aleatoriamente para cada uma das etapas.

Foi executado um novo teste, agora utilizando a acurácia da etapa de testes como resultado aceito, utilizando a etapa de validação somente para escolha do parâmetro de custo que corresponde ao melhor modelo.

Na etapa de testes, a acurácia diminuiu, como esperado, para 86,79% devido à utilização de amostras totalmente desconhecidas pelo modelo para se testar o mesmo. A matriz de confusão obtida é apresentada na tabela 5.

Tabela 5. Matriz de confusão na etapa de testes utilizando BD em português.

	Raiva	Alegria	Neutro	Tristeza
Raiva	80,0	20,0	0,00	0,00
Alegria	9,09	72,73	18,18	0,00
Neutro	0,00	0,00	100,0	0,00
Tristeza	0,00	0,00	0,00	100,0

É possível observar que as emoções neutro e tristeza obtiveram uma taxa de reconhecimento de 100%, o que é algo surpreendente para um classificador relativamente simples como o SVM.

5.2.3. Experimento Utilizando Banco de Dados em Português

Os testes iniciais utilizando o banco de dados em português foram realizados utilizando os mesmos parâmetros ótimos obtidos para o modelo com o EMO-DB. Este teste inicial foi realizado para comparação com os próximos testes, visando identificar o quanto uma linguagem influencia em um modelo de reconhecimento de emoções. Na etapa de validação, foi alcançado 49,2% de acurácia.

Pelo fato de a taxa de reconhecimento ter sido muito menor do que a esperada, realizamos um balanceamento da quantidade de amostras, o que resultou, na etapa de validação, em uma acurácia de 82,1%.

Seguindo-se o mesmo procedimento executado para o EMO-DB, realizamos o teste de força bruta como descrito na seção 5.4.1 visando encontrar parâmetros ótimos para o novo banco de dados testado, utilizando 60% dos dados para treinamento e 20% para validação. Os parâmetros ótimos encontrados neste teste foram de 100 ms para a janela de curto prazo e de 1200 ms para a janela de médio prazo. Foi encontrada uma acurácia de 84,4% na etapa de validação utilizando 20 como parâmetro de custo C.

Tabela 6. Matriz de confusão na etapa de testes utilizando BD em português.

	Raiva	Alegria	Neutro	Tristeza
Raiva	50,0	37,5	12,5	0,00
Alegria	0,00	100,0	0,00	0,00
Neutro	0,00	20,00	60,0	20,00
Tristeza	20,00	0,00	0,00	80,00

Os outros 20% das amostras foram utilizados para a etapa de testes, onde foi obtida uma acurácia de 70,83%. Observando a matriz de confusão da tabela 6, pode-se

constatar que houve uma precisão de 100% para a emoção alegria e 80% para a emoção tristeza, o que é um bom resultado, considerando que os áudios não receberam um pré-processamento, o que poderia aumentar consideravelmente a precisão obtida.

Na figura 5, é apresentado graficamente uma comparação dos resultados entre os dois bancos de dados testados com a aplicação aqui desenvolvida.

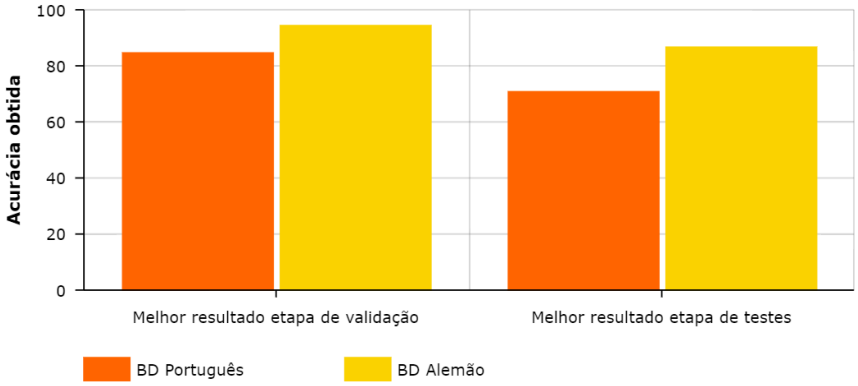


Figura 5. – Comparação dos resultados obtidos com os dois bancos de dados.

7. Conclusões

Através de todos os testes realizados é possível observar a diferença da acurácia final obtida entre os dois banco de dados. A principal razão que encontramos se deve à origem das amostras de áudio. O banco de dados alemão EMO-DB têm uma qualidade muito superior, com mínimos ruídos, enquanto que os áudios em português tem uma qualidade muito inferior, com muitos ruídos e música ao fundo, por exemplo. Portanto, consideramos que o ponto fraco da abordagem utilizando os áudios em português, deve-se à qualidade dos mesmos.

Algumas limitações associadas ao desenvolvimento deste trabalho são: o tamanho do áudio a ser reconhecido, logo, audios longos podem não apresentar taxa condizente com os testes descritos; múltiplas emoções em um mesmo áudio; o regionalismo não foi considerado, podendo causar uma diminuição na taxa de reconhecimento.

Entretanto, a principal limitação do sistema desenvolvido, está diretamente relacionada com esta biblioteca, que foi depender das características de áudios extraídas pela mesma. Talvez, se fossem realizado testes com diferentes características, poderia ser selecionado as melhores e obter um resultado superior.

Como trabalhos futuros, sugerimos uma adição ao modelo desenvolvido neste trabalho, realizando-se um pré- processamento das amostras, o que certamente resultaria em uma melhora na taxa de reconhecimento.

Também seria interessante, realizar a mesma análise deste trabalho com todas as emoções disponíveis no banco de dados de Berlin, afim de comparar os resultados com alguns trabalhos do estado da arte que fizeram o reconhecimento utilizando todas estas emoções.

Uma forma de aumentar a taxa de reconhecimento para este sistema de reconhecimento de emoções, seria realizar uma seleção das características a serem extraídas, otimizando as informações obtidas através das mesmas, o que pode-se esperar que aumentaria razoavelmente a acurácia do sistema.

Referências

ALVA, M. Y.; NACHAMAI; PAULOSE, J. A comprehensive survey on features and methods for speech emotion detection. In: 2015 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT). [S.l.: s.n.], 2015. p. 1–6.

IRIYA, R. Análise de sinais de voz para reconhecimento de emoções. Dissertação (Mestrado) — Curso de Engenharia e Sistemas Eletrônicos, Universidade de São Paulo, 2014.

GIANNAKOPOULOS. Pyaudioanalysis: An open-source python library for audio signal analysis. Public Library of Science One, v. 10, n. 12, dez 2015.

GUNES, H. et al. Emotion representation, analysis and synthesis in continuous space: A survey. In: FG. [S.l.: s.n.], 2011.

HOUWER, J. D.; HERMANS, D. Cognition and Emotion Reviews of Current Research and Theories. New York: Psychology Press, 2010.

LE, D.; PROVOST, E. M. Emotion recognition from spontaneous speech using hidden markov models with deep belief networks. In: 2013 IEEE Workshop on Automatic Speech Recognition and Understanding. [S.l.: s.n.], 2013. p. 216–221.

LUGER, G. F. I. A. 6. ed. São Paulo: Pearson Educação do Brasil, 2013.

PAO, T. long; CHEN, Y. te; YEH, J. heng. Mandarin emotional speech recognition based on svm and nn. 18th International Conference On Pattern Recognition (icpr'06), 2006.

SUJATHA, B.; AMEENA, O. Speech emotion recognition using hmm and gmm and svm models. International Journal Of Professional Engineering Studies (ijpes), p. 311–318, jul 2016.

VERVERIDIS, D.; KOTROPOULOS, C. Emotional speech recognition: Resources, features, and methods. Speech communication, Elsevier, v. 48, n. 9, p. 1162–1181, set 2006.